IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCES IN SIGNAL PROCESSING FOR NON-ORTHOGONAL MULTIPLE ACCESS

Received April 27, 2019, accepted June 10, 2019, date of publication June 19, 2019, date of current version July 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923646

Massively Concurrent Non-Orthogonal Multiple Access for 5G Networks and Beyond

RAZVAN-ANDREI STOICA[1], (Student Member, IEEE), GIUSEPPE THADEU FREITAS DE ABREU[1], (Senior Member, IEEE), TAKANORI HARA[2], (Student Member, IEEE), AND KOJI ISHIBASHI[2], (Member, IEEE)

[1]Focus Area Mobility, Jacobs University Bremen, 28759 Bremen, Germany
[2]Advanced Wireless and Communication Research Center, The University of Electro-Communications, Tokyo 182-8585, Japan

Corresponding author: Razvan-Andrei Stoica (r.stoica@jacobs.university.de)

This work was supported in part by the European Commission in the Framework of the H2020-EUJ-02-2018 Project 5G-Enhance under Grant Agreement 815056, and in part by the Ministry of Internal Affairs and Communications (MIC) of Japan.

ABSTRACT We propose a new signal spreading-based non-orthogonal multiple access (NOMA) schemes that exploits the frame-theoretic design principles to enable the efficient concurrence of multiple users in both downlink (DL) and uplink (UL). In contrast to many other NOMA schemes, the proposed method does not require built-in sparsity on the usage of resources by any individual users, hence referred to as massively concurrent non-orthogonal multiple access (MC-NOMA). Instead, MC-NOMA leverages frames as vectorized ensembles of the individual complex-valued waveform signatures, optimized for low mutual coherence so as to minimize the multi-user interference. A theoretical analysis of the MC-NOMA is offered, which reveals that it can theoretically reach the capacity of the multi-user MIMO channel and outmatches the state of the art in terms of maximum achievable rate over discrete constellations. A detailed description of both the frame-based transmitter design and a newly developed low-complexity efficient decoder is given, which, via simulations, are used to demonstrate that MC-NOMA outperforms the well-established NOMA schemes, such as sparse-coded multiple access (SCMA) and pattern division multiple access (PDMA).

INDEX TERMS Frame theory, massively-concurrent non-orthogonal multiple access, multiuser detection, non-orthogonal communications, tree search decoding.

I. INTRODUCTION

NOMA has emerged in recent years as a promising approach to improve the usage of scarce spectral resources in a seamless manner [1]–[3]. In essence, NOMA schemes aim to provide concurrent access in a true sense, *i.e.*, to simultaneously serve a number of active users larger than the available signal space resources, thus *overloading* the signal space such that no user has exclusive (orthogonal) use of any of the resources.

Non-orthogonality in itself is, however, not a new idea in the field of wireless communications. In early 2000s, for instance, third generation (3G) cellular networks based on code division multiple access (CDMA) applied limited overloading via the utilization of non-orthogonal spreading sequences [4], [5]. And although overloading in CDMA systems did not develop much further due to the raise of orthogonal frequency division multiple access (OFDMA) schemes, adopted in fourth generation (4G) systems for their better spectral efficiency, the approach returned in the form of carrier aggregation mechanisms in 4G long-term evolution (LTE) systems [6], [7].

It can be said, therefore, that in both preceding cases, overloading was used as a late and limited amendment to 3G and 4G systems. The difference with NOMA now is that the non-orthogonal idea is placed at the core of the access architecture, aiming for large scale overloading as a means to achieve higher spectral efficiencies than currently possible [2], [8]–[10].

An important aspect of NOMA is, therefore, the careful design of the overloading structure, and in this regard, NOMA can be qualitatively separated in two categories: power-domain NOMA (PD-NOMA) and code-domain NOMA (CD-NOMA). The PD-NOMA approach relies on the principle of superposition over the power domain.

The associate editor coordinating the review of this manuscript and approving it for publication was Miaowen Wen.

Thereby, users are multiplexed by power coefficients designed to minimize their mutual interference based on the near-far effect and allow for best-possible successive interference cancellation (SIC). Although simple, this idea is known to provide significant gains over equivalent orthogonal multiple access (OMA) systems given that an optimized resource allocation, user selection and clusterization strategy is applied [8], [9], [11]. The drawback of PD-NOMA is, however, that the combinatorial optimization problem associated with the clusterization of users is difficult to solve [12]–[14] limiting the scalability of the approach.

In CD-NOMA, on the other hand, the overloading of non-orthogonally multiplexed users is designed at the bit or symbol level by means of a *code*. This results in a relaxation of the resource allocation complexity, but simultaneously increases the required multiuser detection effort in comparison to the PD-NOMA. Nonetheless, the complexity-performance trade-off between the two leans in favor of CD-NOMA, which is enabled by the application of modern multiuser detection techniques and signal processing [2].

The latter category of CD-NOMA systems can be further divided into subcategories according to the mechanisms used to compress and multiplex user streams onto the existing resources [15]. Specifically, there are CD-NOMA schemes based on *scrambling*, *interleaving*, *spreading* and *coding*.

In scrambling-based CD-NOMA, multiplexation is performed post-modulation by scrambling user streams with low-correlation sequences so as to aid multi-user detection (MUD) via minimum mean squared error (MMSE) with SIC or parallel interference cancellation (PIC) [16], further enhanced by error correction. An example of such an approach is Qualcomm's resource spread multiple access (RSMA) system [17].

Similarly, interleaving- and spreading-based schemes leverage bit/symbol interleaving/spreading techniques in order to alter original user streams, controlling the inter-user interference. One example of interleaving-based NOMA is the interleave division multiple access (IDMA) scheme proposed in [18], in which the low-complexity elementary signal estimator (ESE) algorithm was used for detection, enhanced by extrinsic feedback of the channel decoder structure [19].

In turn, Samsung's interleave-grid multiple access (IGMA) [20], is in fact an enhanced variation of IDMA in which sparse overlapping of interleaved symbols over the available resources is performed similarly to low-density spreading (LDS) [21]–[23], making IGMA an example of the spreading-based approach. Other examples of spreading NOMA are the multi-user shared access (MUSA) [24], the non-orthogonal coded multiple access (NCMA) [25] and the non-orthogonal coded access (NOCA) [26] schemes.

All the aforementioned spreading NOMA approaches rely on randomly or carefully patterned sequences to spread and superimpose streams over all available resources, following the footsteps of classic CDMA, with the distinction that in these modern methods spreading sequences are not binary. The quantized low-correlation random complex spreading sequences utilized in ZTE's MUSA, for instance, aim to provide grant-free support, reduce signaling overhead and improve spurious UL of active users. Simultaneously, a high degree of freedom is maintained via the usage of real and imaginary tertiary sequences $\{-1, 0, 1\}$, so that MMSE-SIC or MMSE-PIC detection can be used at the receiver.

Finally, in code-based CD-NOMA schemes, such as PDMA [27]–[29] and SCMA [30], [31], data streams are mapped into sparse codebooks with particular attributes so as to design their overlap over the different resource domains available, *i.e.* space, time or frequency. Code-based CD-NOMA schemes rely on the message passing algorithm (MPA) algorithm [32] for MUD, such that sparsity, applied to reduce the associated factor graph connectivity [30], is essential for their practical applicability, since the complexity of MPA scales exponentially with the codebook size.

To elaborate more on the role of sparsity in existing CD-NOMA designs, consider again PDMA schemes where certain linear mappings are applied to different resource domains in order to multiplex users onto a group of existing resource elements (REs). As a consequence, the PDMA performance is directly controlled by the pattern matrix defining the resource allocation to multiple users in a differentiable manner [28]. In essence such pattern matrices are nothing but linear binary mappings of active users onto existing resources with or without power control. In turn, in SCMA the bits originating from multiple users are mapped directly to sparse complex spreading codewords that form individual codebook layers, extending the sparse superposition to both time and frequency resources [30]. The performance of SCMA is therefore dictated by the sparse resource mapping policy and by the codebook designs, such that the optimization of the latter is targeted jointly [31]. Finally, multidimensional constellation "shaping" is employed in SCMA, so as to reduce the collisions among active users, especially in the DL case where transmit symbols observe the same channel realizations per each active user.

A. CONTRIBUTIONS OF THIS WORK

As illustrated above, literature on CD-NOMA is very eclectic, spanning over many schemes and their variations. Some recurrent ideas are, however, identifiable across these related works, in light of which this article focuses on answering the following three fundamental questions:

1) CAN CD-NOMA BE ABSTRACTED UNDER A UNIFIED FRAMEWORK?

We formulate in this sense an abstract *compressive mapping* between the collective active user information symbol vector, and respectively, the available REs. At this abstraction level, we identify the last component as a linear allocator of REs. This operation is identified as the *synthesis* linear mapping under the prism of *Frame Theory* [33], [34]. We follow this remark into reducing the CD-NOMA non-orthogonal multiplexing problem to a *frame design formulation*. The latter amounts to assigning optimized signature

waveforms to active users, optimally spreading them over all resources simultaneously. The approach therefore integrates the core ideas of the different NOMA categories previously discussed.

2) WHAT ARE THE OPTIMUM CRITERIA FOR SUCH A FRAMEWORK?

We provide an information-theoretic description of the key design criteria for optimum design of the complex frame that linearly overloads the symbol information vector over the available REs. The analysis yields the main frame-theoretic attributes and features required to simultaneously maximize the system's performance and ease multiuser detection. We prove that the design criteria thus found represent the global optimum in terms of performance across all schemes for the continuous-input-continuous-output memoryless channel (CCMC) case. Analytical results are also provided, which illustrate that the proposed design, referred to as MC-NOMA, consistently outperforms the state of the art over discrete-input-continuous-output memoryless channels (DCMCs).

3) IS THE OPTIMUM PERFORMANCE PRACTICALLY ACHIEVABLE?

In order to demonstrate the practical feasibility of MC-NOMA, a low-complexity multiuser receiver is described, which is shown via simulations to reach the performance of the joint maximum likelihood (ML) detector. Simulation results on the relationship between the complexity and the performance of the MC-NOMA decoder are also given.

B. PAPER OVERVIEW

The paper is organized as follows. In Section II the frame-theoretic abstraction of CD-NOMA systems is introduced. The optimized design criteria of the proposed model, as well as the methods to synthetically generate the desired frames are discussed in Sections III and IV. The MUD tree algorithm is outlined and discussed theoretically in detail in Section V. In Section VI simulations illustrating the performance of MC-NOMA are given and discussed. The article ends with concluding remarks in Section VII.

NOTATION

The following notation will be used throughout. Matrices, column vectors and scalars are denoted $\mathbf{X}$, $\boldsymbol{x}$ and x respectively. The transpose, complex conjugate, and transpose complex conjugate of matrices and vectors are denoted by the superscripts $^{\mathrm{T}}$, * and $^{\mathrm{H}}$, respectively, and the Frobenius, ℓ^2- and ∞-norms are respectively denoted $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$. The inner and Hadamard products are denoted by $\langle\cdot,\cdot\rangle$ and $\cdot\circ\cdot$, respectively. The identity matrix of size M is denoted by $\mathbf{I}_M$. Lastly, the operator diag$(\cdot)$ has a dual purpose: for a vector $\boldsymbol{x}$, it generates the diagonal matrix with diagonal entries given by $\boldsymbol{x}$; and respectively, for a matrix $\mathbf{X}$ it extracts the main entries as a column vector.

II. FRAME-THEORETIC FORMULATION OF CD-NOMA

A. NON-ORTHOGONAL CONCURRENT ACCESS MODEL

A high-level view of the essentials is needed to attain a satisfying abstract framework of CD-NOMA systems. To this extent let us formalize the basic ideas of the CD-NOMA schemes starting from the most fundamental units of information, the active users' bits.

Assume that at any given point a communication system is supposed to serve K active users under the constraint of having available a group of $M < K$ orthogonal REs, *e.g.* frequency subcarriers. Without loss of generality consider that all the K users have the same source bit generation rate leading to them using similar constellations of size Q. As a result, a total of $B = K\log_2(Q)$ bits need to be served simultaneously.

It is clear that such a system is interference limited, and that its decodability does not scale without a bound. Furthermore, a *compressing mapping* must be used to multiplex the information bits to the transmit available resources, such as

$$\mathcal{V} : \{0,1\}^B \to \mathbb{C}^M, \tag{1}$$

with $\mathcal{V}$ decomposable as $\mathcal{F}\circ\mathcal{B}$, *i.e.*

$$\mathcal{B} : \{0,1\}^B \to \mathbb{C}^K, \tag{2a}$$
$$\mathcal{F} : \mathbb{C}^K \to \mathbb{C}^M, \tag{2b}$$

given individual constellation symbol mappings of each independent active user leading to a common *super-symbol* $\boldsymbol{s} \triangleq [s_1, s_2, \ldots, s_K] \in \mathbb{C}^K$ to be multiplexed and radiated over a signal space $\mathbb{C}^M$.

Based on the above, the model of concurrent access over an additive white Gaussian noise (AWGN) channel resumes to the linear system

$$\boldsymbol{y} = \mathbf{P}^{\frac{1}{2}}\mathbf{F}\boldsymbol{s} + \boldsymbol{n}, \tag{3}$$

The complex over-the-air transmitted signal vector

$$\boldsymbol{x} \triangleq \mathbf{P}^{\frac{1}{2}}\mathbf{F}\boldsymbol{s}, \tag{4}$$

becomes therefore a non-orthogonal superposition of *compressed* active users information symbols.

Optional power control strategies could be applied to $\boldsymbol{x}$ in (4) via the diagonal power policy matrix $\mathbf{P}$. Furthermore, *shaping* and *detection decorrelative techniques* may be applied at the level of the symbol constellation mapping $\mathcal{B}$ from (2a). Lastly, the end-level *multiplexing / allocation / spreading* strategy over the available resources is employed by the linear mapping $\mathbf{F}$ corresponding to $\mathcal{F}$ from (2b).

Following (3), $\mathbf{F} \in \mathbb{C}^{M\times K}$ with $M < K$ is a rank-deficient matrix spanning at most $\mathbb{C}^M$. It thus follows by Frame Theory [33], [34] that $\mathbf{F}$ represents the synthesis operator from $\mathbb{C}^K$ to $\mathbb{C}^M$ whose columns are formed by frame vectors over $\mathbb{C}^M$. This general perspective of overloaded NOMA as a Frame Theory problem requires a brief description of some of its fundamentals, which is offered in the next subsection.

B. PRIMER ON FRAME THEORY AND RELATED TOPICS

Frame Theory was pioneered accidentally by Gabor in his seminal work [35], as the filters proposed thereby are in fact the precursors of Gabor frames, which are subsets of general Harmonic frames [34]. Despite this first appearance in the 1940's, a more complete mathematical formalism of Frame Theory has only been developed relatively recently [33].

The general focus of Frame Theory is the study of *overcomplete representations*, also known as *redundant* or *non-orthogonal representations*, *i.e.* generalizations of orthonormal bases. As a consequence, frame-theoretic constructs can be leveraged to analyze and better understand the limits of the generic CD-NOMA model introduced in Section II-A. In the sequel we succinctly introduce a Frame Theory primer sufficient for the scope of this article. The interested reader is referred to [34] and references therein for further details.

A matrix $\mathbf{F} \in \mathbb{C}^{M\times K}$ with $M \leq K$ is said to be a *frame* over the finite Hilbert space $\mathbb{C}^M$ if its linearly dependent column vectors $\boldsymbol{f}_k \in \mathbb{C}^M$ satisfy [34]

$$\alpha \left\|\boldsymbol{y}\right\|_2^2 \leq \sum_{k=1}^{K} \left|\langle \boldsymbol{f}_k, \boldsymbol{y}\rangle\right|^2 \leq \beta \left\|\boldsymbol{y}\right\|_2^2, \quad \forall \boldsymbol{y} \in \mathbb{C}^M, \tag{5}$$

where finite coefficients $0 < \alpha < \beta$ are referred to as the *highest lower*, and the *lowest upper* frame bounds, respectively.

The *analysis matrix* $\mathbf{F}^{\mathrm{H}}$ of a frame performs a *redundant linear transform* which maps a lower dimensional space onto a higher one, *i.e.*

$$\mathbf{F}^{\mathrm{H}} \;:\; \mathbb{C}^M \;\mapsto\; \mathbb{C}^K, \tag{6}$$

and conversely, the *synthesis matrix* compresses higher dimensional information into a lower dimension, *i.e.*

$$\mathbf{F} \;:\; \mathbb{C}^K \;\mapsto\; \mathbb{C}^M. \tag{7}$$

From the perspective of an expansive mapping, such as the one enforced via the analysis matrix, it is natural to define a *frame redundancy* [33], [34],

$$\rho(\mathbf{F}) \triangleq \frac{K}{M}, \tag{8}$$

which measures the extent of the expansion.

On the other hand, from the perspective of the synthesis matrix, the frame redundancy of equation (8) can be seen as *overloading factor*. In the context of CD-NOMA schemes, the overloading factor captures the fact that information of K active users is transmitted over $M < K$ resource units. Hence, for the purposes of this article, we focus on the synthesis aspect of compressed information via overloading.

Obviously, frames generalize classical orthonormal bases. Indeed, any orthonormal basis can be considered a frame with $M = K$, and unit-norm orthogonal column such that $\alpha = \beta = 1$. In this particular scenario, the analysis and synthesis matrices are dual linear unitary transforms capable of decomposing and recovering vectors via their subsequent dual projections. In light of the latter, it is of interest to relate frames and orthonormal basis both qualitatively and quantitatively. In Frame Theory three properties may be identified as significant for this purpose: *tightness*, *incoherence* and *unit-normality*.

Tightness, measured by the *frame potential (FP)* [36],

$$\mathrm{FP}(\mathbf{F}) \triangleq \left\|\mathbf{F}^{\mathrm{H}}\mathbf{F}\right\|_{\mathrm{F}}^2 \geq \frac{K^2}{M}, \tag{9}$$

outlines the energy profile diffused by the overcomplete decomposition over $\mathbb{C}^K$ of all vectors in $\mathbb{C}^M$.

In particular, a frame is said to be tight if $\alpha = \beta$, which reduces equation (5) to

$$\sum_{k=1}^{K} \left|\langle \boldsymbol{f}_k, \boldsymbol{y}\rangle\right|^2 = \alpha \left\|\boldsymbol{y}\right\|_2^2, \quad \forall \boldsymbol{y} \in \mathbb{C}^M, \tag{10}$$

such that a tight frame satisfies a relaxed variation of Parseval's theorem.

A special case of tight frames is when all vectors $\boldsymbol{f}_k$ are unit-norm, which further implicates that $\alpha = \beta = \rho(\mathbf{F})$.

A tight frame is guaranteed [34] to attain the lower bound in (9) as well as to achieve a perfect reconstruction from the synthesis domain to the analysis domain via its *frame operator*, defined as [34],

$$\mathbf{S}_{\mathbf{F}} \;\triangleq\; \mathbf{F}\mathbf{F}^{\mathrm{H}} \overset{\substack{\text{unit-norm}\\ \text{tight}\,\mathbf{F}}}{=} \frac{K}{M}\,\mathbf{I}_M. \tag{11}$$

The last equality in (11), valid for unit-norm tight frames, is an important result in Frame Theory [36]. In general, the spectral properties of the frame operator $\mathbf{S}_{\mathbf{F}}$ mirror the energy profile of the frame, as the lowest and highest eigenvalues of $\mathbf{S}_{\mathbf{F}}$ determine the frame bounds [33].

Incoherence is the frame attribute that describes the spatial angular displacement of the frame vector pairs $(\boldsymbol{f}_k, \boldsymbol{f}_\ell)$, $k \neq \ell$. Quantitatively, incoherence is measured by the *mutual coherence* [34]

$$\mu(\mathbf{F}) \triangleq \max_{k\neq\ell} \frac{\left|\langle \boldsymbol{f}_k, \boldsymbol{f}_\ell\rangle\right|}{\left\|\boldsymbol{f}_k\right\|_2 \cdot \left\|\boldsymbol{f}_\ell\right\|_2} = \max_{k\neq\ell} \left|\cos\left(\measuredangle\left(\boldsymbol{f}_k, \boldsymbol{f}_\ell\right)\right)\right|. \tag{12}$$

In the case of unit-norm column vectors $\mathbf{f}_k$ it follows from (12) that $\mu(\mathbf{F}) = \max_{k<\ell} |g_{\mathbf{F},k\ell}|$, that is, the highest off-diagonal entry of the frame Gram matrix

$$\mathbf{G}_{\mathbf{F}} \triangleq \mathbf{F}^{\mathrm{H}}\mathbf{F}. \tag{13}$$

The squared Frobenius norm of the Gram matrix, $\mathbf{G}_{\mathbf{F}}$, is the *sum of squared correlations (SSC)* of the frame vectors. Its minimization relates therefore to a design criterion often employed in CD-NOMA schemes to optimize the active user multiplexing and inherent system performance, *e.g.* MUSA [24] and PDMA [28], [29]. However, a more important result regarding frame incoherence is the Welch Bound (WB) [37]

$$\mu(\mathbf{F}) \geq \sqrt{\frac{K-M}{M(K-1)}}. \tag{14}$$

In contrast to the FP, the mutual coherence bound is not always achievable [37]. In fact, the only frames achieving the

WB are the *equiangular tight frames (ETF)*, which are the closest equivalent of orthonormal bases in terms of incoherence. The existence of ETF is, however, limited to particular dimension pairs (M, K) [38], [39]. An example of a family of ETFs is the regular simplices over Harmonic structures [38], which exist for all $K = M + 1$, and are obtained by removing the $(M+1)$-th row of a complete unitary $(M+1)$-DFT matrix scaled appropriately by $\frac{1}{\sqrt{M}}$.

Lastly, *unit-normality* is a normalization attribute which ensures fairness in the energy distribution among the column vectors of a frame $\mathbf{F}$, *i.e.* $\|\boldsymbol{f}_k\|_2 = 1, \ \forall k$. The frame-theoretic definitions and notions reviewed above are by no means exhaustive, representing the bare minimum necessary to follow the rest of the article.

C. FRAME-THEORETIC CONSTRUCTION OF MC-NOMA

In light of all the above the abstract massively concurrent access model over the AWGN channel introduced in (3) can now be extended to a more realistic realization over fading communication channels as follows.

1) DOWNLINK

For convenience of exposition, the DL NOMA scenario is addressed first, which happens to be also timely as the DL case is currently gaining traction as a solution for ultra reliable low-latency communications (URLLC) and massive machine-type communications (mMTC) in fifth generation (5G) systems. Concrete applications are vehicle-to-everything (V2X), *e.g.* semaphore-less automated traffic junctions, tele-operated valet parking, conventional autonomous driving dynamic maps & information exchange; massive data dissemination and distributed processing in large-scale fog computing, *e.g.* automated factory / logistics optimization operations applied at the edge networks; and high-throughput data streaming, *e.g.* augmented reality (AR) / virtual reality (VR) museum and exhibition tours, large eGaming events, etc.

In such cases, broadcast information need to be downlinked simultaneously with data streams towards a massive number of individual users, usually also at high-rates. Hence, the general problem considered so far becomes relevant, *i.e.* serving K users in DL at once over $M < K$ units of resources, *e.g OFDMA subcarrier tones, slotted time frames, spatial streams*. Without loss of generality, we continue with the assumption that the resources are M orthogonal tones, such that received signal at the k-th user can be written as

$$\boldsymbol{y}_k = \mathbf{H}_{k,\mathrm{DL}}\mathbf{F} \cdot \boldsymbol{s} + \boldsymbol{n}_k, \tag{15}$$

where $\mathbf{H}_{k,\mathrm{DL}} \triangleq \mathrm{diag}([h_{1,k}, h_{2,k}, \ldots, h_{M,k}]^T) \in \mathbb{C}^{M\times M}$ is a diagonal matrix containing the coefficients of the Rayleigh fading channel between the base station (BS) and k-th user equipment (UE), $\boldsymbol{s} \in \mathcal{S}^K \subset \mathbb{C}^K$ is the vector containing the symbols of *all users*, assumed to be taken from the same Q-ary constellation $\mathcal{S}$, the vector $\boldsymbol{n}_k \sim \mathcal{CN}(0, \sigma\mathbf{I}_M)$ is AWGN, and the columns of the *frame* $\mathbf{F} \in \mathbb{C}^{M\times K}$ are the *modulated signature waveforms* of active users.

2) UPLINK

A *dual formulation* to the above DL applies for the UL NOMA scenario [22], [24], [30]. In this case, K active users transmit simultaneously to a BS such that the received BS signal is described by

$$\boldsymbol{y} = \sum_{k=1}^{K} \mathbf{H}_{k,\mathrm{UL}}\boldsymbol{f}_k \cdot s_k + \boldsymbol{n}, \tag{16a}$$

$$= \sum_{k=1}^{K} \left(\mathrm{diag}\left(\mathbf{H}_{k,\mathrm{UL}}\right) \circ \boldsymbol{f}_k\right) \cdot s_k + \boldsymbol{n}, \tag{16b}$$

$$= \mathbf{H}_{eq,\mathrm{UL}} \cdot \boldsymbol{s} + \boldsymbol{n}, \tag{16c}$$

where $\mathbf{H}_{eq,\mathrm{UL}}$ is an *equivalent frame-precoded multiple access channel* matrix given by

$$\mathbf{H}_{eq,\mathrm{UL}} \triangleq [\mathrm{diag}(\mathbf{H}_{1,\mathrm{UL}}) \circ \boldsymbol{f}_1, \ldots, \mathrm{diag}(\mathbf{H}_{K,\mathrm{UL}}) \circ \boldsymbol{f}_K], \tag{17}$$

with each diagonal matrix $\mathbf{H}_{k,\mathrm{UL}}$ describing the UL channel between the k-th user and the BS.

The duality between the DL and UL MC-NOMA cases can be seen by observing that, also for the DL case, one can define for each user an *equivalent frame-precoded DL channel matrix*

$$\mathbf{H}_{eq,k,\mathrm{DL}} \triangleq \mathbf{H}_{k,\mathrm{DL}} \cdot \mathbf{F} \tag{18}$$
$$= [\mathrm{diag}(\mathbf{H}_{k,\mathrm{DL}}) \circ \boldsymbol{f}_1, \ldots, \mathrm{diag}(\mathbf{H}_{k,\mathrm{DL}}) \circ \boldsymbol{f}_K],$$

such that the DL model is, from a mathematical standpoint, a particular case of the UL one.

III. ACHIEVABLE SUM-RATES AND OPTIMIZATION CRITERIA OF MC-NOMA

As described in the previous section, the MC-NOMA approach applies both to DL and UL, having at its core the design of the "*spreading*" frame $\mathbf{F}$. In this section, we shall seek information-theoretical criteria to design $\mathbf{F}$ under the assumption that CSI at the transmitter (CSIT) is not available, so as not to add latency nor hinder scalability, and thus retain the ability of the scheme to satisfy URLLC and mMTC requirements, fundamental to 5G systems.

We therefore focus on a sum-rate analysis of the proposed MC-NOMA scheme assuming, for simplicity but without loss of generality, that the multi-dimensional symbol $\boldsymbol{s} \in \mathcal{S}^K$ multiplexing the signals from all K active users is constrained to a unit average power. Furthermore, the AWGN receiver thermal noise is assumed to be spectrally flat with the same statistics across all receivers and energy density of N_0.

A. ACHIEVABLE DOWNLINK SUM-RATE WITHOUT CSIT

1) CONTINUOUS INPUT CONTINUOUS OUTPUT MEMORYLESS CHANNEL

Under the assumptions outlined above, in the CCMC case the continuous-input signal is such that $\boldsymbol{s} \sim \mathcal{CN}(0, \frac{1}{K}\mathbf{I}_K)$, and in the DL scenario it is considered that each user is capable of performing MUD to detect its own signal s_k in the optimal

joint ML sense. The ergodic sum-rate capacity across all resources at any given k-th user is therefore given by

$$C_{k,\mathrm{DL}}^{\mathrm{CCMC}} = \mathbb{E}_{\mathbf{H}_k}\left[\log_2 \det\left(\mathbf{I}_M + \underbrace{\frac{1}{KN_0}}_{\mathrm{SNR}} \mathbf{H}_k \overbrace{\mathbf{F}\mathbf{F}^{\mathrm{H}}}^{\text{frame operator } \mathbf{S}_\mathbf{F}} \mathbf{H}_k^{\mathrm{H}}\right)\right], \quad (19)$$

where we have (and shall for the remainder of this subsection), replaced $\mathbf{H}_{k,\mathrm{DL}}$ with $\mathbf{H}_k$ for notational simplicity.

It is clear that both the channel $\mathbf{H}_k$ and the frame $\mathbf{F}$ are factors that greatly influence total throughout. But under the assumption of no CSIT, the design of $\mathbf{F}$ is independent of channel realizations.

Under these conditions, the optimization criterion for the design of $\mathbf{F}$ can be identified as follows. First, noticing that the argument of the determinant in equation (19) is a Hermitian matrix, we may invoke Hadamard's inequality for the determinant of a positive-semidefinite square matrix $\mathbf{A}$

$$\det \mathbf{A} \leq \prod_{k=1}^{K} a_{kk}, \quad \forall \mathbf{A} \in \mathbb{C}^K, \quad (20)$$

with equality attained only for diagonal matrices [40, p. 233].

Since the channel matrix $\mathbf{H}_k$ is diagonal, it follows that in order to maximize the sum-rate in equation (19), the frame operator $\mathbf{S}_\mathbf{F}$ must be diagonal, which in turn implicates the requirement that $\mathbf{F}$ is *tight*. In addition, due to the unavailability of CSIT, and under the assumption of ergodicity across resources, it is found that $\mathbf{F}$ must be *equi-normal*. Without loss of generality, letting the norm of all norm vectors $\mathbf{f}_k$ be the unit, it is concluded that the optimal frame $\mathbf{F}$ is an unit-norm tight frame (UNTF), thus satisfying equation (11), which reduces equation (19) to

$$C_{k,\mathrm{DL}}^{\mathrm{CCMC}} = \mathbb{E}_{\mathbf{H}_k}\left[\log_2 \det\left(\mathbf{I}_M + \frac{1}{MN_0}\mathbf{H}_k\mathbf{H}_k^{\mathrm{H}}\right)\right]. \quad (21)$$

The above expression evaluates furthermore in the case of the ergodic Rayleigh channel $\mathbf{H}_k$ to a closed analytical form given by [41, Th. 2] as

$$C_{k,\mathrm{DL},Rayl.}^{\mathrm{CCMC}} = \int_0^\infty \log_2\left(1 + \frac{1}{MN_0}\lambda\right)\lambda^{K-M}\exp^{-\lambda} \cdot \sum_{m=0}^{M-1} \frac{m!\cdot\left[L_m^{K-M}(\lambda)\right]^2}{(m+K-M)!}\, d\lambda, \quad (22)$$

where λ represents a marginalized singular value dimension of the equivalent non-negative definite matrix $\mathbf{H}_k\mathbf{H}_k^{\mathrm{H}}$ and L_m^j are the (m, j)-associated Laguerre polynomials [41].

We remark that thanks to the generality of the formulation introduced in Subsection II-A, a similar analysis leading to the expression in equation (19) applies to a multitude of other NOMA schemes, including *e.g.* SCMA, PDMA and MUSA. Unlike MC-NOMA, however, in those schemes the multiplexation mapping of user information onto resources is based on intricate designs which embed sparsity, limit entries and/or introduce redundancy. As a result, for those schemes, the term in equation (19) corresponding to the frame operator $\mathbf{S}_\mathbf{F}$ does not reduce (in general) to the form in equation (11), which not only complicates the analysis of their maximum achievable sum-rates, but also implies the *loss of degrees-of-freedom (DoF)*.

In comparison, MC-NOMA leverages all DoF available through the usage of dense complex frames $\mathbf{F}$. In fact, by force of the Hadamard inequality of equation (20), it follows that the CCMC capacity derived in equation (21) is the optimum achievable ergodic sum-rate of any DL NOMA scheme without CSIT. In other words, MC-NOMA has the theoretical potential to outperform all other NOMA schemes at DL in terms of maximum achievable sum-rate. More importantly, such potential can in fact be realized in practice, requiring to that end only that:

a) for any pair (M, K), a frame $\mathbf{F}$ can designed so that its frame operator satisfies equation (11); and
b) that a decoder enabling each user to detect its own signal s_k in the optimal joint ML sense can be designed.

In the sequel, both a frame design algorithm and an efficient decoder will be described, which satisfy these conditions and thus demonstrate that MC-NOMA is not only theoretically but also *de-facto* superior to existing NOMA schemes such as PDMA [27] and SCMA [30], [31].

2) DISCRETE INPUT CONTINUOUS OUTPUT MEMORYLESS CHANNEL

We obtain in this subsection the DL sum-rate at any given user k in the DCMC case, in analogous fashion and under the same conditions as above. By definition, this is given by the maximum mutual information

$$R_{k,\mathrm{DL}}^{\mathrm{DCMC}} \triangleq \max_{p(\boldsymbol{s})\in\mathcal{S}^K} I(\boldsymbol{s};\boldsymbol{y}_k) \quad (23a)$$

$$= \max_{p(\boldsymbol{s})\in\mathcal{S}^K} \left(B - H(\boldsymbol{s}|\boldsymbol{y}_k)\right), \quad (23b)$$

where we emphasize that for simplicity it is assumed that the symbols of the K users in $\boldsymbol{s}$ are sampled out of the same Q-ary constellation $\mathcal{S}$, such that $\boldsymbol{s} \in \mathcal{S}^K$.

In addition $B \triangleq K\log_2(Q)$ is used in (23) to denote the maximum cumulative entropy at the source. The expression in (23) can be further reduced to

$$R_{k,\mathrm{DL}}^{\mathrm{DCMC}} = B - \sum_{i=1}^{2^B}\int_{\boldsymbol{y}_k} p(\boldsymbol{s}_i,\boldsymbol{y}_k)\log_2\left(\frac{1}{p(\boldsymbol{s}_i|\boldsymbol{y}_k)}\right)d\boldsymbol{y}_k \quad (24a)$$

$$= B - \sum_{i=1}^{2^B}\int_{\boldsymbol{y}_k} p(\boldsymbol{s}_i)p(\boldsymbol{y}_k|\boldsymbol{s}_i) \quad (24b)$$

$$\cdot \log_2\left(\frac{p(\boldsymbol{y}_k)}{p(\boldsymbol{s}_i)p(\boldsymbol{y}_k|\boldsymbol{s}_i)}\right)d\boldsymbol{y}_k. \quad (24c)$$

The mutual information in (23) is known to be maximized by an uniform input discrete distribution [42], *i.e.* $p(\boldsymbol{s}) = \frac{1}{2^B}$. What is more, given the communication model in (18) it

follows that

$$p(\boldsymbol{y}_k|\boldsymbol{s}_i) = p\left(\mathbf{H}_k\mathbf{F}\boldsymbol{s}_i + \boldsymbol{n}_k|\mathbf{H}_k\mathbf{F}\boldsymbol{s}_i\right) \tag{25a}$$

$$= (\pi N_0)^{-M} \exp\left(-\frac{\|\overbrace{\boldsymbol{y}_k - \mathbf{H}_k\mathbf{F}\boldsymbol{s}_i}^{\text{noise } \boldsymbol{n}_k}\|_2^2}{N_0}\right) \tag{25b}$$

$$= (\pi N_0)^{-M} \exp\left(-\frac{\|\boldsymbol{n}_k\|_2^2}{N_0}\right) = p(\boldsymbol{n}_k). \tag{25c}$$

It is thus clear that the conditional distribution of the output given the input under fixed channel conditions $\mathbf{H}_k$ and a non-orthogonal combining rule $\mathbf{F}$ is determined by the underlying noise distribution. In addition, using Bayes' rule and marginalization, the probability density function of $\boldsymbol{y}$ can be derived further from (25) as

$$p(\mathbf{y}_k) = \sum_{\ell=1}^{2^B} p(\boldsymbol{s}_\ell)p(\mathbf{y}_k|\boldsymbol{s}_\ell) \tag{26a}$$

$$= \frac{1}{2^B}\sum_{\ell=1}^{2^B} \frac{1}{(\pi N_0)^M} \exp\left(-\frac{\|\boldsymbol{y}_k - \mathbf{H}_k\mathbf{F}\boldsymbol{s}_\ell\|_2^2}{N_0}\right) \tag{26b}$$

$$= \frac{1}{2^B}\sum_{\ell=1}^{2^B} p(\mathbf{H}_k\mathbf{F}\boldsymbol{s}_\ell + \boldsymbol{n}_k|\mathbf{H}_k\mathbf{F}\boldsymbol{s}_\ell)\,. \tag{26c}$$

In light of the latter, equation (23) becomes

$$R_{k,\mathrm{DL}}^{\mathrm{DCMC}} = B - \frac{1}{2^B}\sum_{i=1}^{2^B}\int_{\boldsymbol{n}_k} p(\boldsymbol{n}_k) \cdot \log_2\left[\frac{\sum_{\ell=1}^{2^B} p(\mathbf{H}_k\mathbf{F}\boldsymbol{s}_\ell+\boldsymbol{n}_k|\mathbf{H}_k\mathbf{F}\boldsymbol{s}_i)}{p(\boldsymbol{n}_k)}\right] d\boldsymbol{y}_k \tag{27a}$$

$$= B - \frac{1}{2^B}\sum_{i=1}^{2^B}\mathbb{E}_{\boldsymbol{n}_k}\left[\log_2\sum_{\ell=1}^{2^B}\exp(\eta_{k,\mathrm{DL}}(\ell,i))\right], \tag{27b}$$

where for convenience we implicitly defined

$$\eta_{k,\mathrm{DL}}(\ell,i) \triangleq \frac{\|\boldsymbol{n}_k\|^2 - \|\mathbf{H}_k\mathbf{F}(\boldsymbol{s}_\ell - \boldsymbol{s}_i) + \boldsymbol{n}_k\|^2}{N_0}. \tag{28}$$

Taking the expectation of the expression in equation (27) over realizations of $\mathbf{H}_k$ and $\boldsymbol{n}_k$, and assuming $\mathbf{F}$ is optimized so as to satisfy equation (11), we obtain the ergodic sum-rate in the DCMC case, *i.e.*

$$C_{k,\mathrm{DL}}^{\mathrm{DCMC}} = \mathbb{E}_{\mathbf{H}_k,\boldsymbol{n}_k}\left[R_{k,\mathrm{DL}}^{\mathrm{DCMC}}\right] \tag{29}$$

$$= B - \frac{1}{2^B}\sum_{i=1}^{2^B}\mathbb{E}_{\mathbf{H}_k,\boldsymbol{n}_k}\left[\log_2\sum_{\ell=1}^{2^B}\exp(\eta_{\mathrm{DL}}(\ell,i))\right].$$

From equations (21) and (29), it can be seen that in fact the MC-NOMA system is equivalent to a multiple-input multiple-output (MIMO) overcomplete transmission scheme with more transmit antennas than receive antennas [43], [44, Chap. 10]. In particular, equation (21) is similar to the capacity of MIMO systems [41], and equation (29) is similar to the achievable sum-rate of MIMO systems with discrete input [45], [46].

Lastly, although already made evident by the notation itself, let us emphasize that the expressions in equations (21) and (29) are the achievable sum-rates of a k-th user overall M resources. The system's total achievable DL sum-rates, under the assumption of users with equivalent SNR conditions are, for the CCMC and DCMC cases, respectively given by

$$C_{\mathrm{DL}}^{\mathrm{CCMC}} = K \cdot C_{k,\mathrm{DL}}^{\mathrm{CCMC}}, \tag{30a}$$

$$C_{\mathrm{DL}}^{\mathrm{DCMC}} = K \cdot C_{k,\mathrm{DL}}^{\mathrm{DCMC}}. \tag{30b}$$

B. ACHIEVABLE UPLINK SUM-RATE WITHOUT CSIT

Next, let us analyze the ergodic sum-rate system capacity of UL MC-NOMA under the same conditions as above, considering that the k-th user employs the vector $\boldsymbol{f}_k$, *i.e.*, the k-th column of the frame $\mathbf{F}$, in order to spread its signal across the M resources.

1) CONTINUOUS INPUT CONTINUOUS OUTPUT MEMORYLESS CHANNEL

In light of the duality established by equations (17) and (18), the ergodic achievable rate of UL MC-NOMA in the CCMC is straightforwardly given by

$$C_{\mathrm{UL}}^{\mathrm{CCMC}} = \mathbb{E}_{\mathbf{H}_{eq,\mathrm{UL}}}\left[\log_2\det\left(\mathbf{I}_M + \frac{1}{KN_0}\mathbf{H}_{eq,\mathrm{UL}}\mathbf{H}_{eq,\mathrm{UL}}^{\mathrm{H}}\right)\right]. \tag{31}$$

We remark that $\mathbf{H}_{eq,\mathrm{UL}} \in \mathbb{C}^{M\times K}$ is, under the definition in equation (5), a frame of the same size as its embedded component $\mathbf{F}$.

However, due to the structure imposed by equation (17), and under the condition of no CSIT, no strategy to design $\mathbf{F}$ can be extracted from equation (31) itself, since there are no guarantees that the properties built into the construction of $\mathbf{F}$ carry over to $\mathbf{H}_{eq,\mathrm{UL}}$ with random and unknown $\mathbf{H}_{k,\mathrm{UL}}$. Nevertheless, from an overall system design viewpoint, the following arguments can be offered in favor of users utilizing at UL the same spreading vectors used by the BS at DL.

Firstly, overhead is minimized by having the k-th user employ the same vector $\mathbf{f}_k$ both to de-spread the signal intended to it at DL, and to spread its own signal at uplink. Secondly, by having each user spread its signal via a distinct vector $\boldsymbol{f}_k$ of an *incoherent frame* $\mathbf{F}$, not only is mutual interference (MUI) statistically minimized but also is the power of all users spread uniformly across all resources, maximizing the utilization of all DoFs under the lack of CSIT.

2) DISCRETE INPUT CONTINUOUS OUTPUT MEMORYLESS CHANNEL

Finally, again due to the structural equivalence between $\mathbf{H}_{eq,\mathrm{UL}}$ and $\mathbf{H}_{eq,k,\mathrm{DL}}$ described by equations (17) and (18), the ergodic achievable rate of UL MC-NOMA in the DCMC case is derived analogously to its DL counterpart, yielding

$$C_{\mathrm{UL}}^{\mathrm{DCMC}} = B - \frac{1}{2^B}\sum_{i=1}^{2^B}\mathbb{E}_{\mathbf{H}_{eq,\mathrm{UL}},\boldsymbol{n}}\left[\log_2\sum_{\ell=1}^{2^B}\exp(\eta_{\mathrm{UL}}(\ell,i))\right], \tag{32}$$

where

$$\eta_{\mathrm{UL}}(\ell, i) \triangleq \frac{\|\boldsymbol{n}\|^2 - \|\mathbf{H}_{eq,\mathrm{UL}}(\boldsymbol{s}_\ell - \boldsymbol{s}_i) + \boldsymbol{n}\|^2}{N_0}. \tag{33}$$

From both results above, we conclude that the same criteria for the optimum design of $\mathbf{F}$ in the DL – namely, that $\mathbf{F}$ must be a *tight*, *low-coherence* and *equi-normal* complex-valued frame – remain valid for the UL under imperfect CSIT. In the next section we describe how such frames can be constructed for arbitrary $K \geq M$.

IV. OPTIMIZED MC-NOMA TRANSMITTER: LOW-COHERENCE COMPLEX-VALUED FRAME DESIGN

The matter of constructing *tight, low-coherence, equi-normal* frames is a problem of general interest in Frame Theory and the most desirable set of solutions to this problem is represented by ETFs, such that [34], [37]

$$\frac{|\langle \boldsymbol{f}_k, \boldsymbol{f}_\ell \rangle|}{\|\boldsymbol{f}_k\|_2 \cdot \|\boldsymbol{f}_\ell\|_2} = \sqrt{\frac{K-M}{M(K-1)}}, \quad \forall\, k \neq \ell. \tag{34}$$

In other words, in an ETF, all pairs of distinct frame vectors have the same and minimum normalized inner product theoretically possible. Unfortunately, ETFs do not exist for arbitrary (M, K), and therefore in many cases solutions are sought amongst the next class of desirable tight, low-coherence and equi-normal frames which are represented by the set of unit-norm *Grassmannian* frames [37], which satisfy

$$|\langle \boldsymbol{f}_k, \boldsymbol{f}_\ell \rangle| \geq \sqrt{\tfrac{K-M}{M(K-1)}}, \quad \forall\, k \neq \ell, \tag{35}$$

with $\|\mathbf{f}_k\|_2 = 1$, $\forall k$ and so that $\nexists\, \mathbf{A} \in \mathbb{C}^{M\times K}$ with $\mu(\mathbf{A}) < \mu(\mathbf{F})$.

In plain words, the construction of Grassmannian frames relaxes the requirement of achieving the Welch bound in favor of the more practical condition of achieving the lowest possible frame coherence amongst all frames in the same space. As a result, Grassmanian frames in comparison to ETFs always exist for any given dimensionality [37] and act as a superset for prospective ETFs.

Systematic and explicit methods to construct Grassmannian frames do exist [37], [39], [47], but again, also such methods are restrictive in the choice of parameters (M, K). This limitation led to significant effort made in recent years to devise algorithmic, rather than analytical, methods for the generation of low-coherent frames [48]–[53]. And to that end, convex optimization has proven especially useful, for their combined efficacy and flexibility.

In [51]–[53], for instance, a method to construct incoherent frames was described, which relies on the successive and iterative decorrelation of the column vectors of an initial random frame $\mathbf{F}$, constrained to an unit-norm M-ball enclosing the feasible search regions of each vector. This approach led to the complex successive iterative decorrelation by convex optimization (CSIDCO) algorithm, which is fundamentally described by the problem [53]

$$\left.\begin{array}{l} \underset{\boldsymbol{f}_k \in \mathbb{C}^M}{\text{minimize}}\ \|\tilde{\mathbf{F}}_k^{\mathrm{H}} \boldsymbol{f}_k\|_\infty, \\ \text{subject to}\ \|\boldsymbol{f}_k - \tilde{\boldsymbol{f}}_k\|_2^2 \leq T_k, \end{array}\right\} \quad \forall\, k, \tag{36}$$

where $\tilde{\mathbf{F}}_k$ denotes the matrix resulting from removing the k-th column vector $\boldsymbol{f}_k$ from the frame $\mathbf{F}$, and the radius T_k of the M-dimensional search ball associated with the update of each frame vector $\boldsymbol{f}_k$ is iteratively designed so as to prevent the prospective solution $\boldsymbol{f}_k$ from being collinear with any other $\tilde{\boldsymbol{f}}_\ell$, $\ell \neq k$, *i.e.*

$$T_k \leq 1 - \max_{\ell;\, \ell \neq k} |\langle \boldsymbol{f}_k, \boldsymbol{f}_\ell \rangle|^2. \tag{37}$$

The generically formulated CSIDCO approach described by equation (36) can be explicitly reformulated as the real-valued quadratic program

$$\underset{\boldsymbol{x} \triangleq [\boldsymbol{f}_k; t_{\mathcal{R}}; t_{\mathcal{I}}] \in \mathbb{R}^{2M+2}}{\text{minimize}} \quad \boldsymbol{x}^{\mathrm{T}} \Sigma \boldsymbol{x}, \tag{38a}$$

$$\text{subject to}\ \mathbf{A}_{\mathcal{R},1}\boldsymbol{x} \leq 0,\ \mathbf{A}_{\mathcal{R},2}\boldsymbol{x} \leq 0, \tag{38b}$$

$$\mathbf{A}_{\mathcal{I},1}\boldsymbol{x} \leq 0,\ \mathbf{A}_{\mathcal{I},2}\boldsymbol{x} \leq 0, \tag{38c}$$

$$\boldsymbol{x}^{\mathrm{T}}\mathbf{B}\boldsymbol{x} - 2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} + 1 - T_k \leq 0, \tag{38d}$$

where

$$\Sigma \triangleq \begin{bmatrix} \mathbf{0}_{2M} & \mathbf{0}_{2M\times 2} \\ \mathbf{0}_{2\times 2M} & \mathbf{I}_2 \end{bmatrix} \in \mathbb{R}^{(2M+2)\times(2M+2)}, \tag{38e}$$

$$\mathbf{A}_{\mathcal{R},1} \triangleq \begin{bmatrix} \tilde{\mathbf{F}}_k^{\mathrm{T}} & -\mathbf{1}_{(K-1)\times 1} & \mathbf{0}_{(K-1)\times 1} \end{bmatrix} \in \mathbb{R}^{(K-1)\times(2M+2)}, \tag{38f}$$

$$\mathbf{A}_{\mathcal{R},2} \triangleq \begin{bmatrix} -\tilde{\mathbf{F}}_k^{\mathrm{T}} & -\mathbf{1}_{(K-1)\times 1} & \mathbf{0}_{(K-1)\times 1} \end{bmatrix} \in \mathbb{R}^{(K-1)\times(2M+2)}, \tag{38g}$$

$$\mathbf{A}_{\mathcal{I},1} \triangleq \begin{bmatrix} \tilde{\mathbf{F}}_k^{\mathrm{T}}\mathbf{D}_M & \mathbf{0}_{(K-1)\times 1} & -\mathbf{1}_{(K-1)\times 1} \end{bmatrix} \in \mathbb{R}^{(K-1)\times(2M+2)}, \tag{38h}$$

$$\mathbf{A}_{\mathcal{I},2} \triangleq \begin{bmatrix} -\tilde{\mathbf{F}}_k^{\mathrm{T}}\mathbf{D}_M & \mathbf{0}_{(K-1)\times 1} & -\mathbf{1}_{(K-1)\times 1} \end{bmatrix} \in \mathbb{R}^{(K-1)\times(2M+2)}, \tag{38i}$$

$$\mathbf{D}_M \triangleq \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \otimes \mathbf{I}_M \in \mathbb{R}^{(2M+2)\times(2M+2)} \tag{38j}$$

$$\mathbf{B} \triangleq \begin{bmatrix} \mathbf{I}_{2M} & \mathbf{0}_{2M\times 2} \\ \mathbf{0}_{2\times 2M} & \mathbf{0}_{2\times 2} \end{bmatrix} \in \mathbb{R}^{(2M+2)\times(2M+2)}, \tag{38k}$$

$$\boldsymbol{b}^{\mathrm{T}} \triangleq \begin{bmatrix} \tilde{\boldsymbol{f}}_k^{\mathrm{T}} & 0 & 0 \end{bmatrix} \in \mathbb{R}^{2M+2}. \tag{38l}$$

We refer to our reformulation of CSIDCO given in equation (38) as the quadratic CSIDCO (QCSIDCO) algorithm and remark that it can be easily coded using standard Matlab™ functions, unlike the problem of equation (36).

The final complex solution from equation (38) is obtained as $\mathbf{F}_{\mathrm{Q}} = [1\ \ 1j] \otimes \mathbf{I}_M \mathbf{F}_{\mathrm{Q}}$, recombining the real and imaginary components. The frame $\mathbf{F}_{\mathrm{Q}}$ is a *low-coherent frame* by construction, and unit-normality can be enforced by normalization of the vectors $\mathbf{f}_k$, $\forall k$ obtained after each iteration. However, there is no guarantee that a tight frame is obtained.

In order to mitigate this problem, the frame obtained from equation (38) is further processed via a tightening step given by the polar decomposition [48, Th. 2], which can be described by the optimization problem

$$\min_{\mathbf{F} \in \mathbb{C}^{M\times K},\ \mathrm{UNTF}} \|\mathbf{F}_{\mathrm{Q}} - \mathbf{F}\|_{\mathrm{F}}, \tag{39}$$

whose analytic solution is given respectively by [48]

$$\mathbf{F} = \sqrt{\frac{K}{M}} \left(\mathbf{F}_{\mathrm{Q}}\mathbf{F}_{\mathrm{Q}}^{\mathrm{H}}\right)^{-\frac{1}{2}} \mathbf{F}_{\mathrm{Q}}, \tag{40}$$

where the normalization factor $\sqrt{K/M}$ is applied to ensure the unit-normality constraint.

Resuming all the above, the optimized MC-NOMA transmitter under the assumptions of perfect CSI at the receiver (CSIR) and no CSIT, is given by a complex low-coherence UNTF, which can be obtained as described in Algorithm 1.

Algorithm 1 LOW-COHERENT UNTF DESIGN

Inputs: Number of resources M, users K and iterations I
Output: Low-coherent UNTF $\mathbf{F} \in \mathbb{C}^{M\times K}$

1: Generate initial random matrix: $\mathbf{F}_0 \in \mathbb{C}^{M\times K}$
2: Normalize all column vectors of $\mathbf{F}_0$
3: Iteratively update $\mathbf{F}_0$ into $\mathbf{F}_{\mathrm{Q}}$ by solving equation (38), with T_k updated via equation (37), $\forall\, k$ at each iteration
4: Tighten $\mathbf{F}_{\mathrm{Q}}$ obtaining a low-coherent UNTF via the polar factorization of equation (40).

V. OPTIMIZED MC-NOMA RECEIVER: HIGH-PERFORMANCE LOW-COMPLEXITY MUD

Having demonstrated the theoretical potential of MC-NOMA and described how to optimally design its transmitter in Sections III and IV, it is left for us to turn our attention to the receiver side and describe how MC-NOMA signals can be effectively decoupled, with a particular interest in a receiver that can achieve a performance close to that of the ML joint detector, but with at a feasible complexity.

To this end, let us start with the ML receiver, described by

$$\hat{\boldsymbol{s}}_{\mathrm{ML}} = \underset{\boldsymbol{s}\in\mathcal{S}^K}{\operatorname{argmin}} \quad \|\boldsymbol{y} - \mathbf{H}\boldsymbol{s}\|_2^2, \tag{41}$$

where in the above and hereafter, we simplify the notation so that $\mathbf{H}$ can be either interpreted as to denote $\mathbf{H}_{k,\mathrm{DL}}\mathbf{F}$ in the context of DL reception under the model of equation (15), or $\mathbf{H}_{eq,\mathrm{UL}}$ for the UL case as per equation (16).

Notice that the optimum ML detector described in (41) is also equivalent to the joint maximum a posteriori (MAP) MUD under the assumption that the input symbols are sampled independently and uniformly out of the same constellation $\mathcal{S}$. In other words, we hereby diverge from current CD-NOMA literature – which typically favors rather suboptimal receiver techniques such as the MMSE-SIC, MMSE-PIC, MPA, ESE [15] – by seeking a low-complexity detector capable of achieving the near optimum decoding performance.

This is possible in the case of MC-NOMA because of its massively concurrent approach, in which the MC-NOMA receive signal is similar to that of a straightforward overloaded MIMO system. Based on the later analogy the concept of sphere decoding (SD) [54]–[57] can be applied, such that we may relax the problem in equation (41) to

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}\in\mathcal{S}^K;\mathcal{R}_{\mathbf{H}}^2\le\beta^2}{\operatorname{argmin}} \quad \|\boldsymbol{y} - \mathbf{H}\boldsymbol{s}\|_2^2, \tag{42}$$

where the search radius of the system, determined by the linear transform $\mathcal{R}_{\mathbf{H}} \triangleq \|\boldsymbol{y} - \mathbf{H}\boldsymbol{s}\|_2$, is nothing but a maximum allowed residual of the objective at the solution.

It is clear therefore that (42) is in fact equivalent to the ML search problem particularly for $\beta \to \infty$.

The usual procedure of solving (42) is based on the QR decomposition of $\mathbf{H}$ leading to the equivalent problem

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}\in\mathcal{S}^K;\mathcal{R}_{\mathbf{H}}^2\le\beta^2}{\operatorname{argmin}} \quad \|\boldsymbol{y} - \mathbf{Q}\mathbf{R}\boldsymbol{s}\|_2^2 \tag{43a}$$

$$\underset{\boldsymbol{s}\in\mathcal{S}^K;\mathcal{R}_{\mathbf{R}}^2\le\beta^2}{\operatorname{argmin}} \quad \|\underbrace{\mathbf{Q}^{\mathrm{H}}\boldsymbol{y}}_{\tilde{\boldsymbol{y}}\ proj.} - \mathbf{R}\boldsymbol{s}\|_2^2 \tag{43b}$$

$$\underset{\boldsymbol{s}\in\mathcal{S}^K;\mathcal{R}_{\mathbf{R}}^2\le\beta^2}{\operatorname{argmin}} \quad \|\tilde{\boldsymbol{y}} - \mathbf{R}\boldsymbol{s}\|_2^2, \tag{43c}$$

where, due to the fact that the matrix $\mathbf{Q}$ is unitary, $\mathcal{R}_{\mathbf{H}} = \mathcal{R}_{\mathbf{R}}$ such that the residual metric is preserved.

Next, we observe that the overloading rectangular structure of $\mathbf{F}$ imposes an upper trapezoidal structure onto $\mathbf{R}$, *i.e.*

$$\mathbf{R} = \Big[\ \underbrace{\mathbf{R}_1}_{\text{upper triangular}\ \in\mathbb{C}^{M\times M}} \ \Big|\ \overbrace{\mathbf{R}_2}^{\text{dense}\ \in\mathbb{C}^{M\times(K-M)}} \ \Big], \tag{44}$$

which is typical of generalized sphere decoding (GSD), usually applied to underdetermined linear systems [58]–[65].

In conventional GSD, the search starts with the estimation of the $K - M + 1$ symbols corresponding to the last row of $\mathbf{R}$ and proceeds by subsequently solving the SD problem for the remaining $M-1$ symbols based on the remaining rows of $\mathbf{R}$, all of which is conducted subjected to a constrained search radius $\mathcal{R}_{\mathbf{R}}^2$. In principle, the search radius is such that

$$\sum_{i=1}^{M} \left|\tilde{y}_i - \sum_{j=i}^{K} r_{ij}s_j\right|^2 \le \beta^2. \tag{45}$$

In practice, however, in order to reduce complexity, the search is conducted symbol-by-symbol so that distinct radii are applied to each symbol. In particular, the search for the first $K - M + 1$ symbols decoded is conducted over the radius

$$\beta_M^2 \triangleq \left|\tilde{y}_M - \sum_{j=M}^{K} r_{Mj}s_j\right|^2 \le \beta^2, \tag{46}$$

and respectively, for $\ell \in \{M-1, M-2, \ldots, 1\}$ over the radii

$$\beta_\ell^2 \triangleq \Big| \tilde{y}_\ell - r_{\ell\ell} s_\ell - \underbrace{\sum_{j=\ell+1}^{K} r_{\ell j} s_j}_{\text{interference due to past choices}} \Big|^2$$

$$\leq \overbrace{\beta^2 - \underbrace{\sum_{i=\ell+1}^{M} \beta_i^2}_{\text{accumulated residual}}}^{\text{shrinking search radius at } \ell\text{-th row}} . \tag{47}$$

Literature indicates that there are three usual strategies to reduce the search complexity of conventional GSD receivers: a) to break the original GSD problem down into a series of determined SD sub-problems [58]–[60]; b) to apply the method of Tikhonov regularization leading to a square regularized SD problem of size $K \times K$ [61], [62]; and c) to enclose the prospective estimates of the overcomplete dimensions into a planar parametric slab aiming at reduced complexity [64], [65]. The first of the aforementioned strategies suffers from high complexity, and the last two, hereafter referred to as *regularized* GSD and *slab decoding*, respectively, have limitations on their scalability to arbitrarily overloaded systems such as MC-NOMA.

In particular, the regularized GSD approach expands the system to a full-rank matrix $\mathbf{R}$ by addition of trivial equations. The effectiveness of this operation requires, however, that every constellation point has unit-power which is only valid for BPSK and 4-QAM. In order to be extended to higher-order modulations, one needs to linearly decompose them into combinations of unit-power ones, which is either not feasible or leads a substantial increase in complexity.

In turn, in slab decoding, the solutions of the $K-M$ underdetermined equations are grouped into a parametric geometric planar slab meant to reduce the exhaustive search complexity to the $Q^{(K-M)}$ search space. Therefore, although applicable to arbitrary constellations, the approach offers little complexity reduction in systems with high overloading ratios such as MC-NOMA.

Since the usual strategies to reduce the complexity of GSD are not suitable to MC-NOMA receivers, we develop our own in the sequel. To this end, we start with the standard GSD algorithm, which as can be understood from equations (43) through (47), amounts to a depth-first search (DFS) constrained by adapted SD radii. And although the worst-case complexity of such a search is of order $\mathcal{O}(Q^K)$ [66], it has been shown that its stochastic complexity is of polynomial order, with an exponent dependent on the signal-to-noise ratio (SNR) [56], and which can be further reduced by the application of *reduction*, *enumeration* and *pruning* techniques. We therefore focus below on reduction, enumeration and probabilistic pruning schemes suitable to MC-NOMA receivers.

A. LATTICE-BASED REDUCTION

A well-known reduction mechanism that is general enough to be applicable to MC-NOMA is the Lenstra-Lenstra-Lovász (LLL) lattice reduction [67], [68] over infinite lattices. If standard quadrature amplitude modulation (QAM) constellations are used, search recursions can also be accelerated by various methods such as those discussed in [58]–[63], [69] so long as the complex based values in equation (43) are mapped into corresponding real numbered equivalents. Fortunately that is generally possible for standard QAM, and the structure of equation (43c) can be preserved after such mapping by defining,

$$\tilde{\boldsymbol{y}} \xrightarrow{\mathbb{R}} \tilde{\boldsymbol{y}} \triangleq \begin{bmatrix} \mathrm{Re}(\tilde{\boldsymbol{y}}) \\ \mathrm{Im}(\tilde{\boldsymbol{y}}) \end{bmatrix} \in \mathbb{R}^{2M}, \tag{48}$$

$$\mathbf{R} \xrightarrow{\mathbb{R}} \mathbf{R} \triangleq \left[\begin{array}{c|c} \mathrm{Re}(\mathbf{R}) & -\mathrm{Im}(\mathbf{R}) \\ \hline \mathrm{Im}(\mathbf{R}) & \mathrm{Re}(\mathbf{R}) \end{array}\right] \in \mathbb{R}^{2M \times 2K}, \tag{49}$$

$$\boldsymbol{s} \xrightarrow{\mathbb{R}} \boldsymbol{s} \triangleq \begin{bmatrix} \mathrm{Re}(\boldsymbol{s}) \\ \mathrm{Im}(\boldsymbol{s}) \end{bmatrix} \in \mathbb{R}^{2K}, \tag{50}$$

with the new search space given by the projection of the individual $\mathcal{S}$ constellations on the real and imaginary axes.

It is worth noting that without loss of generality in the case of QAM complex constellations this operation translates to obtaining integer point search lines across the real and imaginary axes which will hereafter be denoted as $\mathcal{Z}$, such that the original problem in $\mathcal{S}^K$ reduces to an *underdetermined discrete least squares* problem over the search space $\mathcal{Z}^{2K}$.

Finally, since such a complex-to-real transformation preserves the 2-norm, the residual metrics of inequalities (45)-(47) remain valid within the subsequent dimensionality doubling of the problem, *i.e.* $(2M, 2K)$ instead of (M, K).

B. SCHNORR-EUCHNER ENUMERATION

A very important aspect in the structure of GSD as a DFS tree search algorithm is represented by the order in which the potential solutions are evaluated.

The right strategy to enumerate the set of prospective solutions for each user symbol is therefore valuable to reduce complexity of the search [55], [56], because once a candidate solution is achieved at the bottom of the tree with a lower residual than the previous, the search radius can be reduced for the next symbol.

The update is based on reducing the sphere radius provided that the residual of the currently found solution is lower than the one of the older solution which was in fact used to set the existing search radius.

A well-known and effective enumeration technique is the Schnorr-Euchner (SE) enumeration proposed in [70]. This method leads to a zig-zag search over the available feasible solution space and is also referred to as the Schnorr-Euchner sphere decoding (SESD) [71]. Applying SE ordering over each level, the search tree itself is sorted, arranging the solutions in accordance with non-decreasing residuals on the left-most branches, which are the ones that are visited

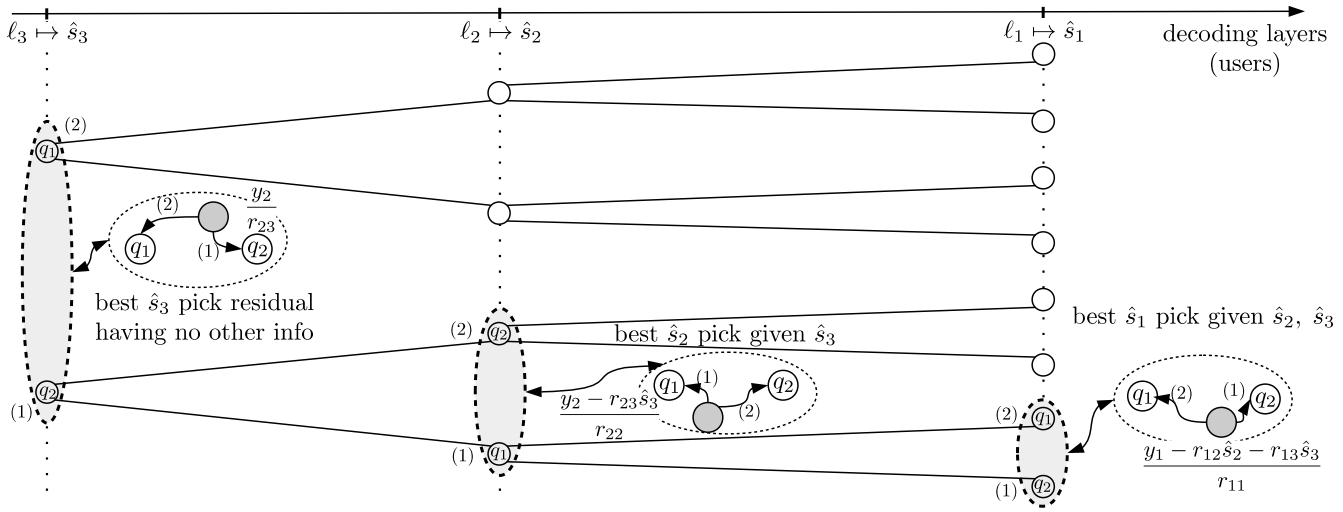


FIGURE 1. Illustration of SE ordering and tree search optimization for a (2, 3) demonstrative overcomplete access system serving 3 users over 2 resources with binary bipolar constellations.

first, as depicted in Figure 1. This strategy, coupled with the search radius update described earlier, effectively increases the performance, leading (in average) to significantly lower complexity polynomial exponents [56].

C. PROPOSED TREE PRUNING

Unfortunately, the LLL reduction and SE enumeration methods discussed above are not sufficient to curb the complexity of MC-NOMA receivers, especially in the case of large overloading ratios, because the detection of the first $K - M + 1$ symbols is governed by the last row of the trapezoidal matrix $\mathbf{R}$ described in equation (44), and conducted under a single constraint radius as described by equation (46).

In other words, tree pruning cannot be performed before reaching the M-th detection layer in the tree, because only at that stage the corresponding M-th system equation can be utilized, given its associated radius inequality (46), to perform tree pruning and reduce complexity.

However, by revisiting the incremental inequalities

$$\beta_k^2 \triangleq \Big| \underbrace{\Big(\tilde{y}_M - \sum_{j=k+1}^{K} r_{Mj} s_j \Big)}_{\text{past residual}} - \underbrace{r_{Mk} s_k}_{k-\text{th pick}} \Big|^2 \leq \beta^2 \tag{51a}$$

$$= |r_{Mk}|^2 \Big| \Big(\frac{\tilde{y}_M - \sum_{j=k+1}^{K} r_{Mj} s_j}{r_{Mk}} \Big) - s_k \Big|^2 \leq \beta^2 \tag{51b}$$

for all the detection layers $k \in \{M, M+1, \ldots, K\}$ strong insights into further search optimizations in the form of tree pruning can be derived, which are pursued in the sequel.

1) ALGEBRAIC REORDERING

In GSD problems, the conventional DFS tree search of SD is expanded over a family of trees rooted by the possible constellation points of the top detection layer [62], [63].

It is therefore desirable to perform subtree cuts as early as possible in order to maintain a low-decoding complexity. Since the first opportunity to prune the tree is at the M-th layer, as detailed above, our focus is shifted to the M-th row equation of the linear system associated with equation (43c).

It can be seen from the incremental inequalities (51) that the entries of the $M-$th row of the matrix $\mathbf{R}$ act as weights for the possible detected symbols, which affect the residual contribution some symbol selections impact onto others.

As a consequence, in order to perform reduction of the search space over the matrix $\mathbf{R}$, it is of interest to penalize bad choices more at higher layers than at lower layers. In other words, the power of the weights should be sorted in non-decreasing fashion from layer M to layer K, such that

$$|r_{MM}|^2 \leq \cdots \leq |r_{Mk}|^2 \leq \cdots \leq |r_{MK}|^2. \tag{52}$$

It is also desirable for the $|r_{Mk}|^2$ entries to have the highest possible magnitudes given the matrix $\mathbf{R}$. Since satisfying such a requirement would lead to a combinatorial problem, we relax this criterion to a maximum sum-of-squares requirement on the last row entries of the trapezoidal matrix $\mathbf{R}$, *i.e.*

$$\max_{\mathbf{R}} \sum_{k=M}^{K} |r_{Mk}|^2. \tag{53}$$

The reduction conditions (52) and (53) imply a column and row reordering of the matrix $\mathbf{R}$, or alternatively, of the matrix $\mathbf{H}$, as well as of its associated vectors $\boldsymbol{y}$ and $\boldsymbol{s}$. Mathematically these reorders resume to system permutations along the two axes of the linear system, *i.e.*

$$\underbrace{\mathbf{P}_r \boldsymbol{y}}_{\text{permuted } \boldsymbol{y}} = \underbrace{\mathbf{P}_r \mathbf{H} \mathbf{P}_c}_{\text{permuted } \mathbf{H}} \cdot \underbrace{\mathbf{P}_c^{\mathrm{H}} \boldsymbol{s}}_{\text{permuted } \boldsymbol{s}}. \tag{54}$$

It follows that the optimized algebraic reduction problem of sorting (52), and respectively of finding the maximum

sum of squares (54), reduces to solving the optimum row and column permutations $\mathbf{P}_r$ and $\mathbf{P}_c$, respectively. To this end the linear system in (54) is reinterpreted given the QR decomposition as

$$\boldsymbol{y}_{\mathrm{P}} = \overbrace{\mathbf{Q}\mathbf{R}}^{\mathbf{P}_r\mathbf{H}\mathbf{P}_c} \cdot \boldsymbol{s}_{\mathrm{P}} \tag{55a}$$

$$\mathbf{Q}^{\mathrm{H}}\boldsymbol{y}_{\mathrm{P}} = \mathbf{R}\boldsymbol{s}_{\mathrm{P}} \tag{55b}$$

$$\tilde{\boldsymbol{y}}_{\mathrm{P}} = \mathbf{R}\boldsymbol{s}_{\mathrm{P}} \tag{55c}$$

where $\boldsymbol{y}_{\mathrm{P}} \triangleq \mathbf{P}_r\boldsymbol{y}$ and $\boldsymbol{s}_{\mathrm{P}} \triangleq \mathbf{P}_c^{\mathrm{H}}\boldsymbol{s}$ are the permuted versions of $\boldsymbol{y}$ and $\boldsymbol{s}$, respectively.

In plain words, as a step preceding the tree pruning algorithm to be described in the sequel, it is advantageous to carefully design the QR decomposition of (55), with the aim to maximize the objective in (53), while satisfying the constraints in (52).

The idea leverages the fact that the QR decomposition preserves the column norm of the input matrix and performs column sorting based on the non-decreasing 2-norm criterion. Indeed, similar column reordering strategies have been studied before as a means to enhance detection performance of MIMO systems and reduce the decoding complexity of space-time codes [63], [72]–[75].

Following preceding literature, the permutations are performed so as to force the rows of $\mathbf{R}$ to accumulate the highest possible sum of squares after canceling each sub-diagonal column matrix elements during the QR decomposition.

Algorithm 2 RUBIK-QR DECOMPOSITION

Inputs: $\mathbf{H} \in \mathbb{C}^{M\times K}$, $M \leq K$
Outputs: unitary $\mathbf{Q} \in \mathbb{C}^{M\times M}$, trapezoidal $\mathbf{R} \in \mathbb{C}^{M\times K}$, rows permutation $\mathbf{P}_r \in \mathbb{Z}^{M\times M}$, columns permutation $\mathbf{P}_c \in \mathbb{Z}^{K\times K}$

1: Initialize $\mathbf{P}_r = \mathbf{I}_M$, $\mathbf{P}_c = \mathbf{I}_K$, $\mathbf{R} = \mathbf{H}$
2: Sort columns of $\mathbf{R}$ according to rule $\left\|\mathbf{R}_{(:,1)}\right\|_2 \leq \left\|\mathbf{R}_{(:,2)}\right\|_2 \leq \cdots \leq \left\|\mathbf{R}_{(:,K)}\right\|_2$ and obtain columns permutation $\boldsymbol{p}_c \in \mathbb{Z}^K$
3: Update $\mathbf{P}_c \mapsto \mathbf{P}_{c(:,p_c)}$, $\mathbf{R} \mapsto \mathbf{R}\mathbf{P}_c$
4: Sort rows of $\mathbf{R}$ according to rule $\left\|\mathbf{R}_{(1,:)}\right\|_2 \leq \left\|\mathbf{R}_{(2,:)}\right\|_2 \leq \cdots \leq \left\|\mathbf{R}_{(K,:)}\right\|_2$ and obtain rows permutation $\boldsymbol{p}_r \in \mathbb{Z}^M$
5: Update $\mathbf{P}_r \mapsto \mathbf{P}_{r(p_r,:)}$, $\mathbf{R} \mapsto \mathbf{P}_r\mathbf{R}$
6: **for** $c = 1{:}M$ **do**
7: Given $\mathbf{R}_{(:,c)}$, generate Householder reflector $\boldsymbol{u}$[76]
8: Cache column reflector $\mathbf{U}_{(c:M,c)} = \boldsymbol{u}$
9: Apply Householder transform (HHT) to submatrix $\mathbf{R}_{(c:M,c:K)} \mapsto \mathrm{HHT}\left(\boldsymbol{u}, \mathbf{R}_{(c:M,c:K)}\right)$, [76]
10: **end for**
11: Sort last $K-M$ columns by $|\mathbf{R}_{(M,M)}| \leq |\mathbf{R}_{(M,M+1)}| \leq \ldots \leq |\mathbf{R}_{(M,K)}|$ and update permutation $\boldsymbol{p}_c \in \mathbb{Z}^{K-M+1}$
12: Update $\mathbf{P}_{c(:,M:K)} \mapsto \mathbf{P}_{c(:,\boldsymbol{p}_c)}$
13: Retrieve $\mathbf{Q} \mapsto \mathrm{HHT}\left(\mathbf{U}, \mathbf{I}_M\right)$ from cached reflectors

This is implemented efficiently by algebraic reflections given by the Householder Transformation [76], such that the objective in (53) is maximized in polynomial time. After the optimum row permutation is achieved, the M to K columns are permuted based on the M-th row of $\mathbf{R}$ so as to satisfy the non-decreasing criterion of (52).

Since the succession of column and row permutations of $\mathbf{R}$ resemble the block rotations performed to solve a Rubik cube, we refer to the optimized QR decomposition described above as the *Rubik-QR decomposition*. A complete description of the Rubik-QR algorithm used to solve the reduction strategy discussed above is thus offered[1] in Algorithm 2.

2) PROBABILISTIC TREE PRUNING

Even after the application of all the techniques described above, the detection of highly overloaded systems may still be too complex as a result of the impossibility of early tree pruning. In order to cap the complexity of such worst cases, it is desirable to perform early branch cuts based on a suitable threshold. This issue motivated work on various stochastic strategies to identify such thresholds [78]–[80], which rely on noise statistics to generate parametric pruning decisions and therefore are referred to as *probabilistic tree pruning*.

A general framework for probabilistic tree pruning was proposed in [80] encompassing different generic pruning strategies such as uniform pruning, K-best, thresholding and hybrids thereof. Amongst other contributions, a central result on pruning by hard thresholding which links branch cuts to tractable parameters stemming from sound noise and signal space probabilistic distributions was described in [80, Prop. 2]. And although conventional QR decomposition of square matrices was assumed in [80], it is obvious given the unitary transforms involved that the Rubik-QR approach described above preserves the chi-squared distribution of squared residuals, such that the same technique applies to our problem.

In the context of MC-NOMA detection, pruning via a threshold δ at the (earliest possible) M-th detection layer is achieved by following the main result in [80, Eq. (14)], with δ given by

$$\delta = 2\sigma^2 \cdot \gamma^{-1}\Big(1 - \big(1 + \tfrac{d_{min}^2}{4\sigma^2}\big)^{-M}\Big), \tag{56}$$

where d_{min} denotes the minimum distance of the finite constellations transmitted and $\gamma(x)$ is the lower incomplete gamma function

$$\gamma(x) \triangleq \int_0^x \exp(-t)dt. \tag{57}$$

[1] Octave GNU and MATLAB® [77] notations for submatrix and subvector selection are used for simplicity.

Algorithm 3 RUBIK-PTST

Inputs: $\boldsymbol{y} \in \mathbb{C}^M$, $\mathbf{H} \in \mathbb{C}^{M\times K}$, constellation $\mathcal{S}$, σ^2
Outputs: estimates $\boldsymbol{s} \in \mathcal{S}^K$

Set initial search radius $\beta = +\infty$
Compute threshold δ given (56)
Set layer $\ell = K$
Allocate default candidate symbol sets $\forall \ell$ layers, S_ℓ
Apply Rubik-QR reduction
$(\mathbf{Q}, \mathbf{R}, \mathbf{P}_r, \mathbf{P}_c) = \text{RUBIK} - \text{QR}\,(\mathbf{H})$
Update $\boldsymbol{y} \mapsto \mathbf{Q}^{\text{H}}\mathbf{P}_r\boldsymbol{y}$
while (1) **do**
  if $\ell \geq M$ **then**
    if EMPTY(S_ℓ) **then**
      if $\ell = K$ **then**
        break
      else
        Reset S_ℓ to default ▷ clears visit
        Go up the tree $\ell \mapsto \ell + 1$
        continue
      end if
    else
      if set S_ℓ is not visited yet **then**
        Apply SE to $S_\ell \mapsto \text{SE}\left(\mathcal{S}, \mathbf{R}, \boldsymbol{y}_M, \hat{\boldsymbol{s}}_{(\ell+1:K)}\right)$
        if $\ell = M$ **then**
          for all symbols $S_{\ell,i} \in S_\ell$ **do**
            if not conditions (58) **then**
              REMOVE($S_\ell, S_{\ell,i}$) ▷ prune $S_{\ell,i}$ subtree
            end if
          end for
        end if
      end if
    end if
    Get layer ℓ estimate $\hat{\mathbf{s}}_{(\ell)} \mapsto$ POP_FRONT(S_ℓ)
    Go down the tree $\ell \mapsto \ell - 1$
  else
    Solve inner SD sub-problem
    $\hat{\mathbf{s}}_{(1:M-1)} \mapsto \text{SESDTHRESH}\left(\boldsymbol{y}_{(1:\ell)}, \mathbf{R}_{(1:\ell,1:\ell)}, \delta, \mathcal{S}\right)$
    Compute obtained residual $\beta^2_{\text{err}} = \left\|\boldsymbol{y} - \mathbf{R}\hat{\boldsymbol{s}}\right\|_2^2$
    if $\beta^2_{\text{err}} \leq \beta^2$ **then**
      $\beta \mapsto \beta_{\text{err}}$ ▷ Decrease search radius
    end if
    Update search layer $\ell = M$
  end if
end while

Concretely, at any layer $\ell = \{1, 2, \ldots, M\}$ only the subtrees stemming from leafs whose paths fulfill the conditions

$$\sum_{i=\ell}^{M} \beta_i^2 \leq \beta^2, \tag{58a}$$

$$\beta_\ell^2 \leq \delta, \tag{58b}$$

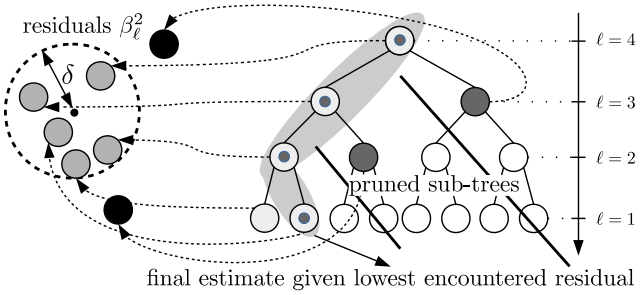


FIGURE 2. Probabilistic tree pruning by δ-thresholding.

are further explored, while the constraint-violating subtrees are cut, as illustrated in Figure 2.

In other words, the canonical cumulated residual constraint (58a) of sphere decoding is complemented by the probabilistic thresholding rule (58b) introduced by leveraging the framework in [80].

We refer to the complete MC-NOMA MUD as the *Rubik probabilistic tree search with thresholding (Rubik-PTST)*, a pseudo-code of which is offered in Algorithm 3. Notice that the inner SESDthresh(·) refers to a standard SESD decoder complemented by the probabilistic tree pruning via thresholding (58b) described earlier. An efficient implementation of such a decoder can be achieved via the fast LLL reduced recursions procedure for finite constellations of [69, Alg.6]. To this end, the problem should be converted to its real equivalent as stated in (48)-(50).

All in all, the proposed Rubik-PTST scheme incorporates the LLL lattice reduction, the Schnorr-Euchner enumeration, the algebraic reordering via the Rubik-QR decomposition, and the optimized threshold-based probabilistic tree pruning schemes into a powerful detector for overloaded systems, whose performance is assessed in the following section.

VI. NUMERICAL RESULTS AND SIMULATIONS

In this section we validate via numerical simulations the models, results and algorithms discussed and proposed in Sections II through V.

A. EFFECTIVENESS OF QCSIDCO FRAME DESIGN

To highlight the transmitter design concepts of MC-NOMA and its inherent advantages some effort is first dedicated to illustrating the structure imposed by the low-coherence tight frames used in MC-NOMA onto the channels perceived by the receivers.

To this end, consider a frame $\mathbf{F}$ obtained via the QCSIDCO method described in Section IV and summarized in Algorithm 1. In MC-NOMA, the transmission frame is designed to minimize the maximum interference among users towards the Welch Bound, leading to an almost uniform coherence level. This is illustrated in Figure 3a, which shows an intensity map of the Gram matrix corresponding to a frame $\mathbf{F} \in \mathbb{C}^{8\times 14}$. This design feature is important to maintain the inter-user interference limited and corresponds to the sparse non-orthogonal multiplexation of users and simultaneous codebook optimization to avoid collapsing of

different symbol combinations onto similar outcomes in the case of SCMA; and to the design of sparse patterns for controlled non-orthogonal combining and interference in the case of PDMA, respectively.

It can be learned from Figure 3a, however, that MC-NOMA achieves the same design objective of the aforementioned state-of-the-art alternatives, *without* requiring sparsity in order to enable the non-orthogonal combining of active users. As a result MC-NOMA leverages all the available resources to the full extent, as opposed to schemes such as SCMA and PDMA, which once again explains the superior achievable rates of MC-NOMA as illustrated in the sequel in Subsection VI-B.

Next, we move to illustrate both the effect of the unknown transmit channel onto the spreading achieved by the optimized low-coherence frame design, and the equivalence between DL and UL in the latter. To this end, the intensity map of an equivalent DL channel $\mathbf{H}_{eq,k,\mathrm{DL}}$ corresponding to the frame $\mathbf{F}$ in Figure 3a and one channel realization is shown in Figure 3b, whereas the intensity map of the UL counterpart $\mathbf{H}_{eq,\mathrm{UL}}$ is shown in Figure 3c.

It can be seen not only that indeed the DL and UL cases are qualitatively similar, but also that the effect of an unknown channel over the MC-NOMA low-coherence frame is (in both DL and UL cases) to slightly increase the overall coherence of the equivalent channel as compared to that of the frame itself, but still in a manner that ensures a gradation of interference powers from various users, much like what is achieved in schemes such as SCMA and PDMA at the expense of larger design complexity and a sparsity requirement which are avoided in the MC-NOMA scheme.

Altogether, Figure 3 illustrates that strongly separable non-orthogonal user signals result from the MC-NOMA frame, regardless of whether the channel is diagonal (as in the DL) or fully rich (as in the UL).

B. ACHIEVABLE RATE PERFORMANCE

To follow with the information theoretic limits, we next compare the achievable DL and UL sum-rates of the proposed MC-NOMA system against those of SCMA [30] and PDMA [27]. These two benchmarks were selected due to their well-known performance gains over other CD-NOMA schemes [15] across both different realistic and idealized scenarios.

For the purpose of the first set of comparisons, we considered a 150% overloaded system with $(M, K) = (4, 6)$, both in DL and UL. The SCMA scheme was optimized via the factor graph [30], [31] corresponding to the sparse matrix

$$\mathbf{F}_{\mathrm{SCMA}} \triangleq \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}. \tag{59}$$

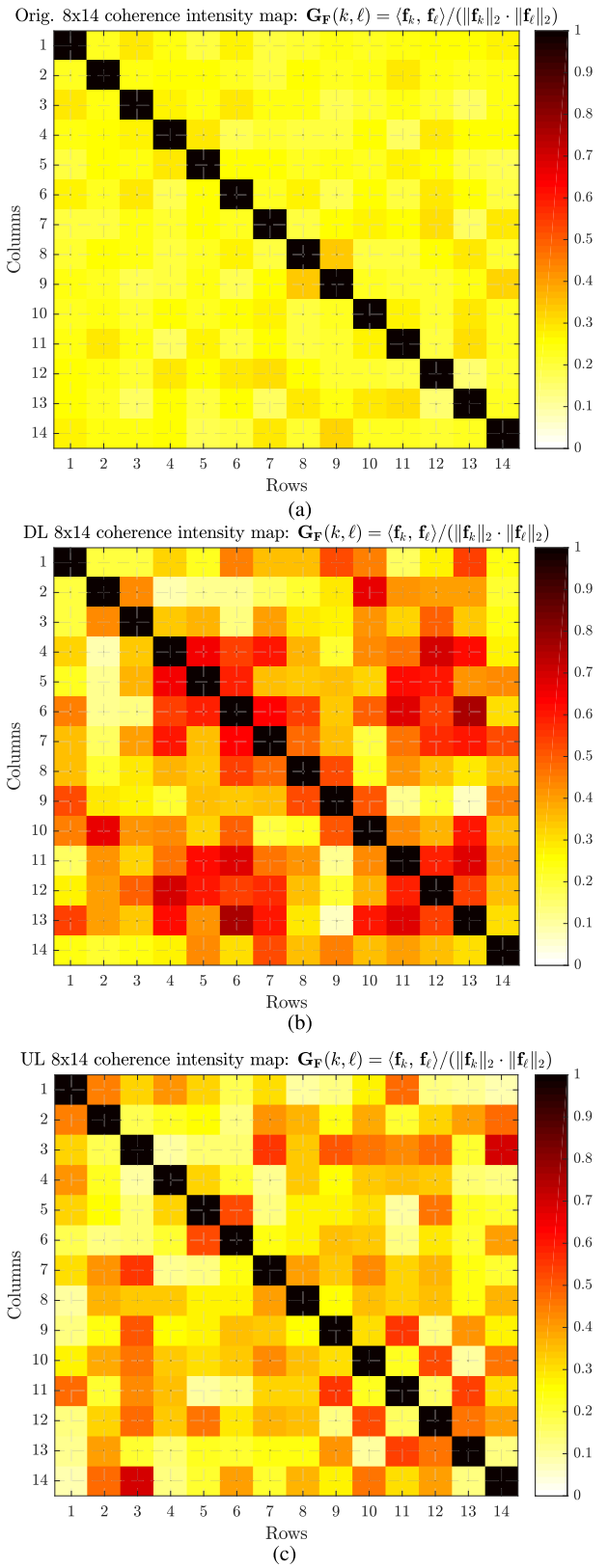


FIGURE 3. Intensity map of Gram matrix $\mathbf{G}_{\mathbf{F}} = \mathbf{F}^{\mathsf{H}}\mathbf{F}$. (a) Incoherent UNTF obtained with Algorithm 1. (b) DL equivalent frame: $\mathbf{H}_{eq,k,\mathrm{DL}}$. (c) UL equivalent frame: $\mathbf{H}_{eq,\mathrm{UL}}$.

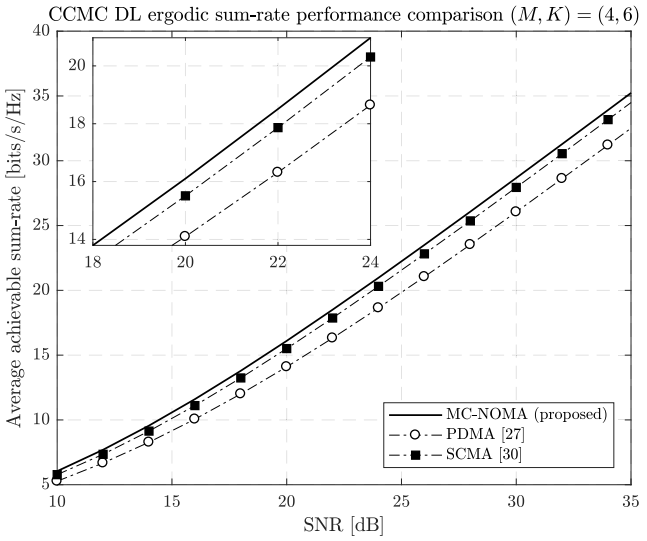


FIGURE 4. Theoretic spectrally normalized CCMC ergodic sum-rate for CD-NOMA DL systems, equation (19). The proposed MC-NOMA curve is equivalent to (21), where the frame F is obtained by Algorithm 1.

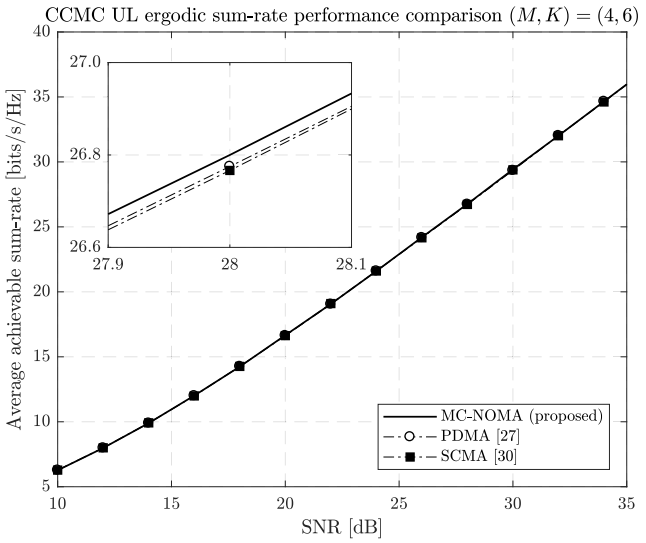


FIGURE 5. Theoretic spectrally normalized CCMC ergodic sum-rate for CD-NOMA UL systems, equation (31). The MC-NOMA curve is given by the proposed frame design F in Algorithm 1.

In turn, the PDMA system was optimized via the pattern multiplexing matrix

$$\mathbf{F}_{\text{PDMA,DL}} \triangleq \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}, \tag{60}$$

for the DL [29] and

$$\mathbf{F}_{\text{PDMA,UL}} \triangleq \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}, \tag{61}$$

for the UL [28], respectively.

The aforementioned multiplexation patterns follow the optimized design criteria across [28]–[30] and aim to provide equivalent comparison conditions for MC-NOMA, SCMA and PDMA, such that in the latter two the non-orthogonal multiplexing of active users is achieved with a maximum overlap of $d_{f,max} = 3$ per resource.

In addition, for the SCMA scheme, the characteristic codebook design principles described in [31] were used throughout the simulations for an optimized baseline. We highlight that in the CCMC scenario these optimization criteria reduce to simple input symbol fixed rotations as described in [30], [31]. The input values were drawn randomly in this case for all schemes from the maximum entropic normalized zero-mean Gaussian distribution. The results are shown in Figures 4 and 5 for the DL and UL scenarios, respectively. It is found that the proposed MC-NOMA system outperforms both SCMA and PDMA in both DL and UL across the SNR range, although the advantage of MC-NOMA over the state- of-the-art benchmarks in the UL case is negligible.

This can be explained by the fact that the design criteria to construct the frames $\mathbf{F}$ employed in the MC-NOMA scheme

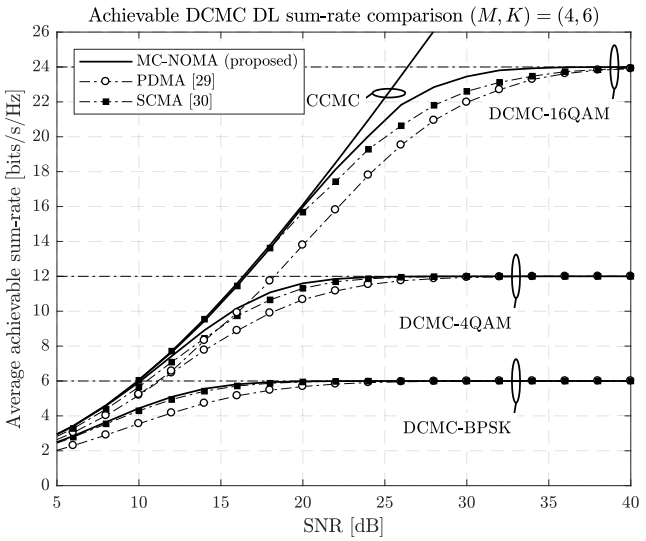


FIGURE 6. DL ergodic sum-rate over the DCMC case across MC-NOMA, SCMA [30], and PDMA [29]. The DL MC-NOMA CCMC sum-rate capacity is plotted as the upper bound of the discrete achievable ergodic sum-rates.

are such that the DL sum-rate of (19) is maximized, while not harming the UL sum-rate as per (31). To rephrase, the low-coherence tight frame construction described in Section IV is more effective and meaningful in shaping the overloaded multiplexation of users at DL than it is at UL, as evidenced by the UL and DL equivalent channels respectively described by (17) and (18). It is therefore to be expected that a similar result is observed in the practical case of systems with discrete-input symbols. This is indeed confirmed in Figures 6 and 7, which are the counterparts of the latter for the DCMC case, and from which it can again be seen that the advantage of MC-NOMA over SCMA and PDMA is more pronounced in DL than in UL. But furthermore,

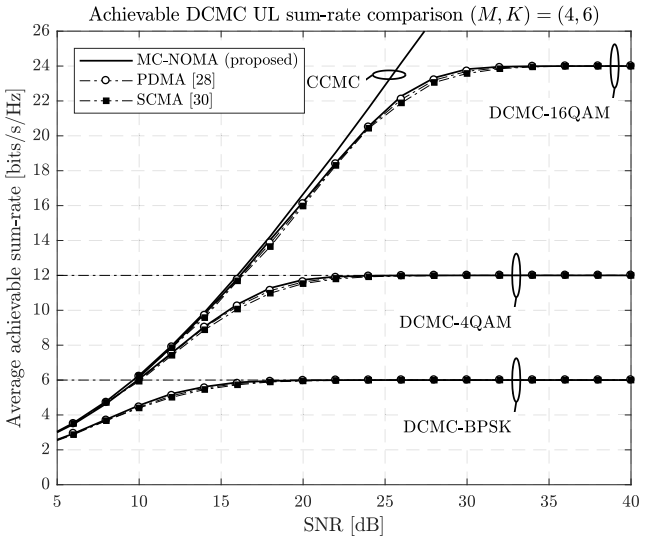


FIGURE 7. UL ergodic sum-rate over the DCMC case across MC-NOMA, SCMA [30], and PDMA [28]. The UL MC-NOMA CCMC sum-rate capacity is plotted as the upper bound of the discrete achievable ergodic sum-rates.

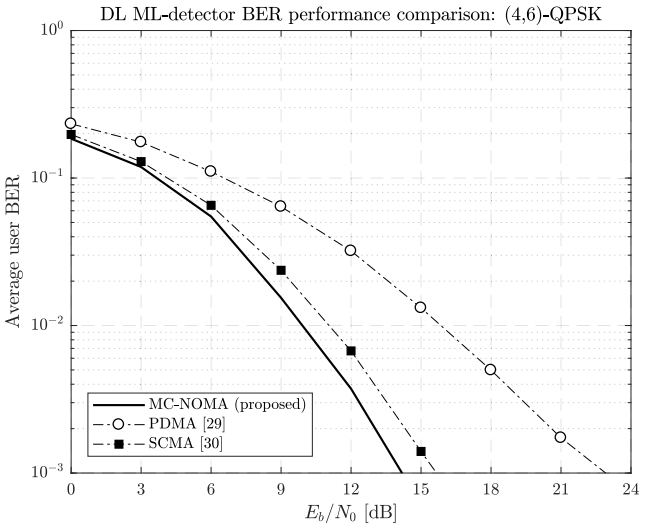


FIGURE 8. Uncoded DL BER performance gain of QPSK MC-NOMA over state-of-the-art SCMA [30], PDMA [29].

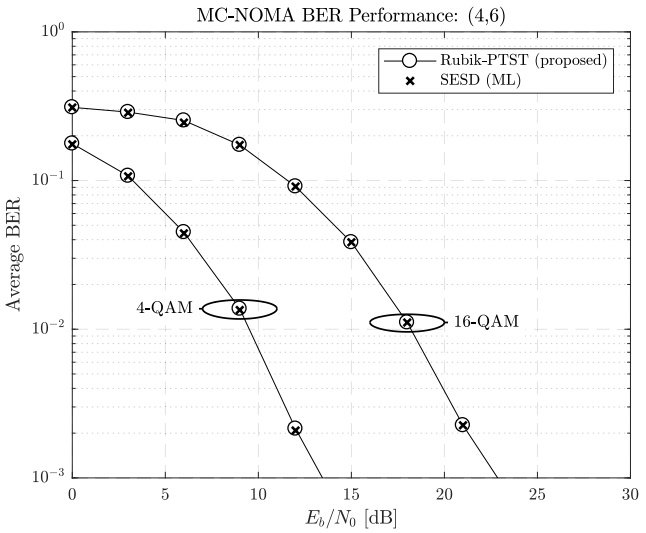


FIGURE 9. BER performance of MC-NOMA with Rubik-PTST and ML detection via SESD [69] optimized GSD in a $(M, K) = (4, 6)$ system with 4-QAM and 16-QAM inputs.

the figures also show that at the knee SNR regions the gap between the proposed MC-NOMA and the state-of-the-art benchmark schemes is larger under DCMC than in CCMC, regardless of the constellation employed, but particularly for constellations of larger order.

This outlines the scalability of the proposed MC-NOMA scheme, which is capable of separating users of an overloaded system by means of complex signature waveforms leading to complex spreading with limited and controlled interference, altogether leading to a more efficient use of the available signal space compared to SCMA and PDMA. Indeed, PDMA uses conventional bits-to-symbols mappings and relies on channel coding in multiplexation onto non-orthogonal and recoverable patterns, while SCMA optimizes the multiplexed non-orthogonal signals avoiding collisions via multi-dimensional signaling and rotations [31], both of which become difficult in larger scales.

Finally to confirm the robustness and practical feasibility of the rate gains MC-NOMA records over the state of the art, Figure 8 highlights the uncoded ML average bit-error rate (BER) for the same $(4, 6)$ system discussed in the prequel. It can be seen therefore that the capacity performance of MC-NOMA over SCMA and PDMA is doubled by an increased MUD capability given the optimized interference management and full exploitation of the available signal space DoFs that MC-NOMA proposed.

C. PERFORMANCE OF Rubik-PTST RECEIVER

As a consequence of the equivalence of MC-NOMA received signals in DL and UL modes, as illustrated in the previous subsection, it is clear that MC-NOMA exhibits similar performance characteristics both at DL and UL.

Motivated by the fact that end users have more stringent complexity restrictions than base stations, and aiming to illustrate the practical feasibility of the MC-NOMA scheme, we choose to address in our simulations the detection performance and complexity of the proposed Rubik-PTST receiver over a DL multicarrier link.

Across all simulations, channels were modeled as independent and identically distributed (iid) Rayleigh fading channels, and the premises and system model described in Section II were followed. In order to obtain each data point in all curves, averages were taken over 40 frame constructions and as many noise and channel realizations as required to ensure sufficient statistical significance.

For the purpose of benchmarking, the performance of our own Rubik-PTST receiver was compared against that of the GSD further optimized by the SESD and accelerated by faster recursion technique proposed in [69], which is known to efficiently reach ML detection performance over finite search sets [80]. Finally, in order to illustrate the ability of

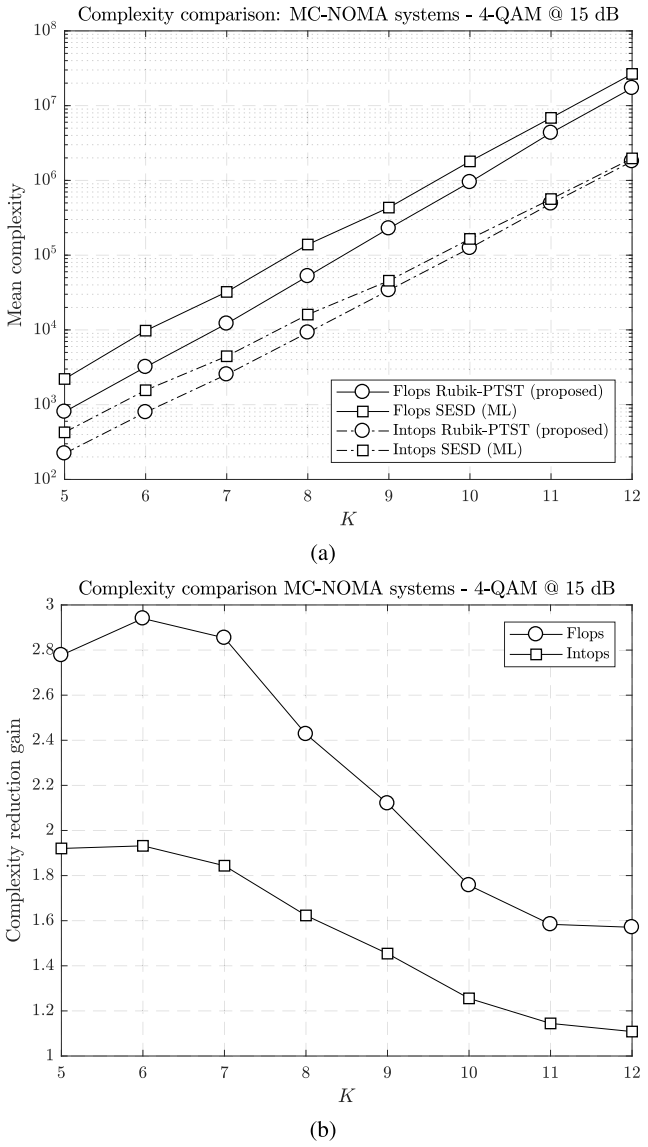


FIGURE 10. Complexity of MC-NOMA systems (M, K) with Rubik-PTST and SESD optimized GSD receivers over 4-QAM at SNR = 15dB, with $M = 4$ and $K = 5$ to 12. (a) Average number of operations. (b) Complexity reduction gain.

the receiver to handle constellations with points of distinct amplitudes and phases, QAM modulation was utilized.

The first set of results is shown in Figure 9 for a system with 150% overloading resulting from $K = 6$ users sharing $M = 4$ resources. Curves for the BER as a function of the energy per bit over noise power (EbN0) corresponding to systems employing 4-QAM and 16-QAM modulation are shown, from which it can be seen that the proposed Rubik-PTST receiver achieves the ML performance, so that altogether it can be understood that that the proposed MC-NOMA scheme with Rubik-PTST detection achieve, in practice, the capacity curves shown in Figures 6 and 7.

In light of the above, a natural question that may arise is what is the complexity associated with the results shown in Figure 9. In particular, it is of relevance to compare the computational cost of the proposed Rubik-PTST receiver

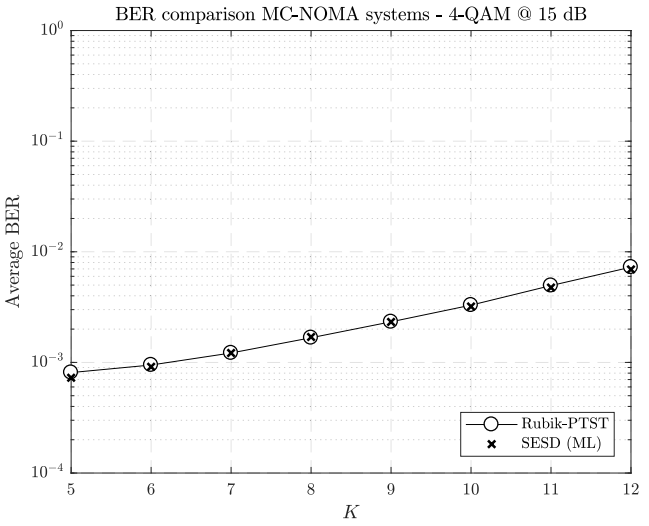


FIGURE 11. Average BER performance of MC-NOMA systems of size (M, K), such that $M = 4$, and respectively, $K \in \{5, 6, 7, 8, 9, 10, 11, 12\}$.

against that of the accelerated and optimized GSD, which is known to be polynomial [56], [69], [80]. Such a comparison is offered in Figure 11, which shows both the absolute and the relative complexities of the two methods as a function of the number of users $K > M$, with the number of resources fixed to $M = 4$, the SNR fixed to 15dB and 4-QAM modulation.

First, absolute numbers are shown in Subfigure 10a, revealing that the complexity of the proposed Rubik-PTST receiver measured in floating point and in integer point operations (flops and intops, respectively) is not only lower than that of state-of-the-art benchmark, but also follows the same trend. We remark that the observed linear growth on K implicates that the complexity order of the Rubik-PTST receiver is polynomial on the overloading ratio K/M.

In addition, the absolute complexity numbers of Subfigure 10a indicate that even at 300% overloading, symbol detection in the Rubik-PTST-decoded MC-NOMA systems simulated thereby is achievable in milliseconds using standard modern computing and without requiring optimized or embedded computation environments, which speaks of the practical feasibility of MC-NOMA altogether.

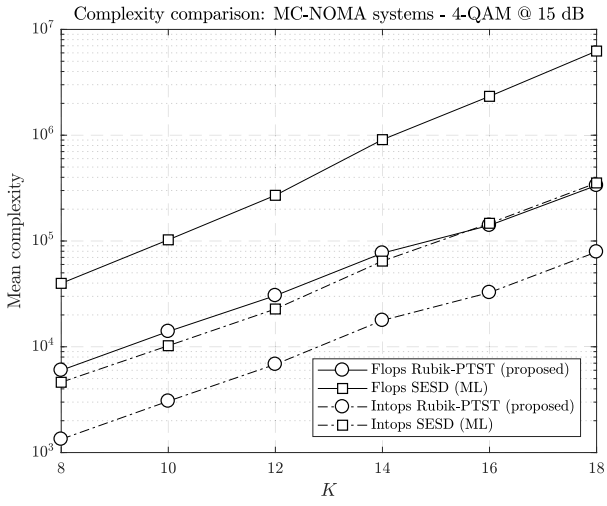
Next, Subfigure 10b shows the relative complexity reduction gain obtained by the Rubik-PTST receiver over the optimized and accelerated GSD [56], [69], [80], defined as

$$G_R \triangleq \frac{\sum_{\mathbf{H}_i;\, i=1}^{I} \text{flops}_{\text{GSD}}}{\sum_{\mathbf{H}_i;\, i=1}^{I} \text{flops}_{\text{Rubik-PTST}}}, \tag{62}$$

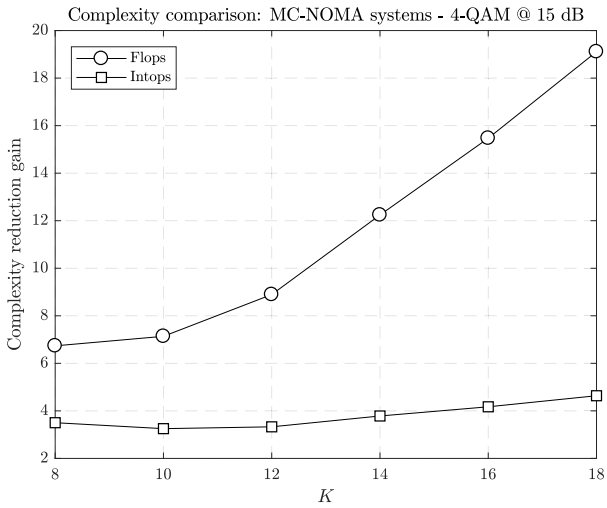
where the number of iterations, *i.e.* I, was set to 1000.

The results reveal that while higher gains are achieved by the Rubik-PTST receiver at moderate overloading factors of up to 200%, the proposed decoder continues to outperform the state-of-the-art also for higher overloading scenarios.

Finally, in order to complement the latter set of results, a comparison of the BER performance of the two decoders for the scenarios studied in Figure 10 is shown in Figure 11,



(a)



(b)

FIGURE 12. Complexity comparison for MC-NOMA systems (M, K), such that $M = \{6, 8, 10, 12, 14, 16\}$, and respectively, $K = M + 2$. (a) Average number of operations. (b) Complexity reduction gain.

which demonstrates that the complexity reduction improvements of Rubik-PTST do not compromise its decoding complexity, which in fact continues to exhibit ML-like behavior.

Having demonstrated the efficiency, efficacy and robustness of MC-NOMA and its Rubik-PTST receiver, we lastly turn our attention to the scalability of the scheme. To this end, we investigate the performance of Rubik-PTST-decoded MC-NOMA systems of size $(M, M + 2)$ with $M \in \{6, 8, 10, 12, 14, 16\}$.

The results are shown in Figure 12, which is a counterpart of Figure 10 for the case of increasing scale with a constant overloading ratio. It can be seen that under such conditions the proposed Rubik-PTST receiver offers larger gain in terms of complexity reduction compared to the state of the art, maintaining the total joint detection time in the order of milliseconds even for systems with $K = 18$ users.

This result can be explained by the impact of the increase in overall resources onto the level of frame coherence attained the optimized frame construction. To elaborate, in light of the WB in (14), under a constant number of resources, the increase in overloading leads to an increase in frame coherence and thus of inter-user interference. On the other hand, under a constant overloading ratio or a constant delta between the number of users and resources, an increase in the absolute number of RE results in higher degrees of freedom. This is efficiently exploited by the QCSIDCO frame optimization scheme, leading to lower inter-user interference. In other words, the MC-NOMA scheme with Rubik-PTST-based detection actually benefits from larger scales, both in terms of performance and relative complexity advantage.

These characteristics indicate that the proposed MC-NOMA scheme and its Rubik-PTST receiver offer a compelling solution both for URLLC and mMTC type applications.

Additionally, these attributes could be enhanced by grant-free protocols leveraging the fact that each user operates over its own signature waveform, potentially leading to even greater latency reduction via the elimination of overhead.

VII. CONCLUSION

We considered problems of non-orthogonal multiplexing in both DL and UL, proposing a new signal spreading-based NOMA scheme which enables the efficient concurrence of multiple users in both DL and UL. To that end, a frame-theoretic formulation of the central NOMA problem was first introduced leading to similar system model expressions for DL and UL CD-NOMA communication systems.

This formulation was then leveraged to design a new NOMA scheme in which frames are utilized as vectorized ensembles of individual complex-valued waveform signatures, optimized for low mutual coherence so as to minimize multi-user interference. In contrast to many other NOMA schemes, the proposed method does not require built-in sparsity on the usage of resources by any individual users, and therefore is referred to as MC-NOMA.

A theoretical analysis of the MC-NOMA was then offered, which revealed that it can theoretically reach the capacity of the multi-user MIMO channel, and outperform state-of-the-art alternatives in terms of maximum achievable rate over discrete constellations. The analysis also revealed the optimum criteria for the design of the MC-NOMA frames, which was then enforced in the QCSIDCO algorithm yielding a detailed description of the frame-based transmitter, later shown via simulations to be indeed effective.

In order to demonstrate also the practical feasibility of the new scheme, an efficient and optimum joint detection scheme for MC-NOMA receivers was proposed in the form of a highly optimized probabilistic pruning tree search algorithm *Rubik-PTST*. This method leverages the particularities of the proposed MC-NOMA scheme to implement a low-complexity, but robust and optimum symbol detector capable of achieving the information theoretic limits previously derived analytically.

Using this receiver, it was then shown via simulations that MC-NOMA outperforms well-established NOMA schemes such as SCMA and PDMA.

REFERENCES

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC-Spring)*, Jun. 2013, pp. 1–5.

[2] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[3] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, Oct. 2017.

[4] H. D. Schotten and H. Hadinejad-Mahram, "Analysis of a CDMA downlink with non-orthogonal spreading sequences for fading channels," in *Proc. IEEE 51st Veh. Technol. Conf. (VTC-Spring)*, vol. 3, May 2000, pp. 1782–1786.

[5] A. Kapur and M. K. Varanasi, "Multiuser detection for overloaded CDMA systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1728–1742, Jul. 2003.

[6] M. Iwamura, K. Etemad, M.-H. Fong, R. Nory, and R. Love, "Carrier aggregation framework in 3GPP LTE-advanced," *IEEE Commun. Mag.*, vol. 48, no. 8, pp. 60–67, Aug. 2010.

[7] C. S. Park, L. Sundström, A. Wallén, and A. Khayrallah, "Carrier aggregation for LTE-advanced: Design challenges of terminals," *IEEE Commun. Mag.*, vol. 51, no. 12, pp. 76–84, Dec. 2016.

[8] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 611–615.

[9] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[10] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C.-L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[11] S. M. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[12] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for MC-NOMA systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[13] Y. Sun, D. W. K. Ng, and R. Schober, "Optimal resource allocation for multicarrier MISO-NOMA systems," in *Proc. IEEE Int. Commun. Conf. (ICC)*, May 2017, pp. 1–7.

[14] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[15] Z. Wu, K. Lu, C. Jiang, and X. Shao, "Comprehensive study and comparison on 5G NOMA schemes," *IEEE Access*, vol. 6, pp. 18511–18519, 2018.

[16] S. Verdú, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[17] Y. Cao, H. Sun, J. Soriaga, and T. Ji, "Resource spread multiple access—A novel transmission scheme for 5G uplink," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–5.

[18] L. Ping, L. Liu, K. Wu, and W. K. Leung, "Interleave division multiple-access," *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 938–947, Apr. 2006.

[19] L. Ping, L. Liu, K. Y. Wu, and L. WK, "On interleave-division multiple-access," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 5, Jun. 2004, pp. 2869–2873.

[20] Q. Xiong, C. Qian, B. Yu, and C. Sun, "Advanced NoMA scheme for 5G cellular network: Interleave-grid multiple access," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–5.

[21] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, Apr. 2008.

[22] J. van de Beek and B. M. Popovic, "Multiple access with low-density signatures," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov./Dec. 2009, pp. 1–6.

[23] R. Hoshyar, R. Razavi, and M. Al-Imari, "LDS-OFDM an efficient multiple access technique," in *Proc. IEEE 71st Veh. Technol. Conf. (VTC-Spring)*, May 2010, pp. 1–5.

[24] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-user shared access for Internet of Things," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–5.

[25] *Considerations on DL/UL Multiple Access for NR*, document R1-162517, 3GPP, LG Electronics, Jan. 2019, pp. 1–4. [Online]. Available: http://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_84b/Docs/R1-162517.zip

[26] *Non-Orthogonal Multiple Access for New Radio*, document R1-165019, 3GPP, Nokia, Jan. 2019, pp. 1–6. [Online]. Available: http://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_85/Docs/R1-165019.zip

[27] J. Zeng, B. Li, X. Su, L. Rong, and R. Xing, "Pattern division multiple access (PDMA) for cellular future radio access," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2015, pp. 1–5.

[28] B. Ren, Y. Wang, X. Dai, K. Niu, and W. Tang, "Pattern matrix design of PDMA for 5G UL applications," *China Commun.*, vol. 13, no. 2, pp. 159–173, Nov. 2016.

[29] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, "Pattern division multiple access—A novel nonorthogonal multiple access for fifth-generation radio networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3185–3196, Apr. 2017.

[30] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 332–336.

[31] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE 80th Veh. Tech. Conf. (VTC-Fall)*, Sep. 2014, pp. 1–5.

[32] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[33] O. Christensen, *An Introduction to Frames and Riesz Bases*, vol. 7. New York, NY, USA: Springer, 2003.

[34] P. G. Casazza and G. Kutyniok, *Finite Frames: Theory and Applications*, 1st ed. Basel, Switzerland: Birkhäuser, 2012.

[35] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Elect. Eng.*, vol. 93, no. 26, pp. 429–441, Jul. 1946.

[36] J. J. Benedetto and M. Fickus, "Finite normalized tight frames," *Adv. Comput. Math.*, vol. 18, no. 2, pp. 357–385, 2003.

[37] T. Strohmer and R. W. Heath, Jr., "Grassmannian frames with applications to coding and communication," *Appl. Comput. Harmon. Anal.*, vol. 14, no. 3, pp. 257–275, May 2003.

[38] M. Fickus and D. G. Mixon, "Tables of the existence of equiangular tight frames," 2015, *arXiv:1504.00253*. [Online]. Available: https://arxiv.org/abs/1504.00253

[39] M. Fickus, J. Jasper, E. J. King, and D. G. Mixon, "Equiangular tight frames that contain regular simplices," *Linear Algebra Appl.*, vol. 555, pp. 98–138, Oct. 2018.

[40] D. J. H. Garling, *Inequalities: A Journey Into Linear Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[41] İ. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999.

[42] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). New York, NY, USA: Wiley, 2006.

[43] D. Tse and P. Viswanath, *Fundamentals Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[44] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, Jun. 2003. doi: 10.1109/JSAC.2003.810294.

[45] C. Xiao, Y. R. Zheng, and Z. Ding, "Globally optimal linear precoders for finite alphabet signals over complex vector Gaussian channels," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3301–3314, Jul. 2011.

[46] S. X. Ng and L. Hanzo, "On the MIMO channel capacity of multidimensional signal sets," *IEEE Trans. Veh. Technol.*, vol. 55, no. 2, pp. 528–536, Mar. 2006.

[47] J. I. Haas and P. G. Casazza, "On the structures of Grassmannian frames," in *Proc. Int. Conf. Sampling Theory Appl. (SampTA)*, Jul. 2017, pp. 377–380.

[48] J. A. Tropp, I. S. Dhillon, R. W. Heath, Jr., and T. Strohmer, "Designing structured tight frames via an alternating projection method," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 188–209, Jan. 2005.

[49] W. Chen, M. R. D. Rodrigues, and I. J. Wassell, "Projection design for statistical compressive sensing: A tight frame based approach," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 2016–2029, Apr. 2013.

[50] B. Dumitrescu, "Designing incoherent frames with only matrix–vector multiplications," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1265–1269, Sep. 2017.

[51] C. Rusu and N. González-Prelcic, "Optimized compressed sensing via incoherent frames designed by convex optimization," 2015, *arXiv:1507.02454*. [Online]. Available: https://arxiv.org/abs/1507.02454

[52] C. Rusu and N. González-Prelcic, "Designing incoherent frames through convex techniques for optimized compressed sensing," *IEEE Trans. Signal Process.*, vol. 64, no. 9, pp. 2334–2344, May 2016.

[53] C. Rusu, N. González-Prelcic, and R. W. Heath, Jr., "Algorithms for the construction of incoherent frames under various design constraints," *Signal Process.*, vol. 152, pp. 363–372, Nov. 2018.

[54] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Math. Comput.*, vol. 44, no. 170, pp. 463–471, Apr. 1985.

[55] H. Vikalo, "Sphere decoding algorithms for digital communications," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 2003.

[56] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. Expected complexity," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2806–2818, Aug. 2005.

[57] H. Vikalo and B. Hassibi, "On the sphere-decoding algorithm II. Generalizations, second-order statistics, and applications to communications," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2819–2834, Aug. 2005.

[58] M. O. Damen, K. Abed-Meraim, and J. C. Belfiore, "Generalised sphere decoder for asymmetrical space-time communication architecture," *Electron. Lett.*, vol. 36, no. 2, pp. 166–167, Jan. 2000.

[59] P. Dayal and M. K. Varanasi, "A fast generalized sphere decoder for optimum decoding of under-determined MIMO systems," in *Proc. Annu. Allerton Conf. Commun. Control Comput.*, 2003, pp. 1216–1225.

[60] Z. Yang, C. Liu, and J. He, "A new approach for fast generalized sphere decoding in MIMO systems," *IEEE Signal Process. Lett.*, vol. 12, no. 1, pp. 41–44, Jan. 2005.

[61] T. Cui and C. Tellambura, "An efficient generalized sphere decoder for rank-deficient MIMO systems," *IEEE Commun. Lett.*, vol. 9, no. 5, pp. 423–425, May 2005.

[62] X.-W. Chang and X. Yang, "An efficient regularization approach for underdetermined MIMO system decoding," in *Proc. Int. Conf. Wireless Commun. Mobile Comput. (IWCMC)*, Aug. 2007, pp. 349–353.

[63] X.-W. Chang and X. Yang, "An efficient tree search decoder with column reordering for underdetermined MIMO systems," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov. 2007, pp. 4375–4379.

[64] K.-K. Wong and A. Paulraj, "Near maximum-likelihood detection with reduced-complexity for multiple-input single-output antenna systems," in *Proc. 38th Asilomar Conf. Signals, Syst. Comput.*, vol. 1, Nov. 2004, pp. 1158–1162.

[65] K.-K. Wong, A. Paulraj, and R. D. March, "Slab-sphere decoding: Efficient maximum-likelihood detection for asymmetric MIMO antenna systems," in *Proc. IEEE 61st Veh. Technol. Conf. VTC-Spring*, vol. 1, May/Jun. 2005, pp. 692–696.

[66] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. Cambridge, MA, USA: MIT Press, 2009.

[67] A. K. Lenstra, H. W. Lenstra, Jr., and L. Lovász, "Factoring polynomials with rational coefficients," *Mathematische Annalen*, vol. 261, no. 4, pp. 515–534, Dec. 1982.

[68] X.-W. Chang, J. Wen, and X. Xie, "Effects of the LLL reduction on the success probability of the Babai point and on the complexity of sphere decoding," *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 4915–4926, Aug. 2013.

[69] A. Ghasemmehdi and E. Agrell, "Faster recursions in sphere decoding," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3530–3536, Jun. 2011.

[70] C. P. Schnorr and M. Euchner, "Lattice basis reduction: Improved practical algorithms and solving subset sum problems," *Math. Program.*, vol. 66, no. 1, pp. 181–199, Aug. 1994.

[71] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.

[72] G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 11, pp. 1841–1852, Nov. 1999.

[73] D. Wübben, R. Böhnke, J. Rinas, V. Kühn, and K. D. Kammeyer, "Efficient algorithm for decoding layered space-time codes," *Electron. Lett.*, vol. 37, no. 22, pp. 1348–1350, Oct. 2001.

[74] D. L. Milliner, E. Zimmermann, J. R. Barry, and G. Fettweis, "A fixed-complexity smart candidate adding algorithm for soft-output MIMO detection," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 6, pp. 1016–1025, Dec. 2009.

[75] R. Y. Chang and W.-H. Chung, "Best-first tree search with probabilistic node ordering for MIMO detection: Generalization and performance-complexity tradeoff," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 780–789, Feb. 2012.

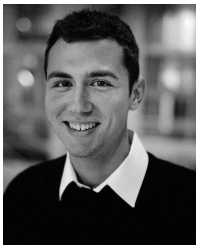
[76] A. S. Householder, "Unitary triangularization of a nonsymmetric matrix," *J. ACM*, vol. 5, no. 4, pp. 339–342, Oct. 1958.

[77] *MATLAB, Version 7.10.0 (R2010a)*, Natick, MA, USA, The MathWorks, 2010.

[78] R. Gowaikar and B. Hassibi, "Statistical pruning for near-maximum likelihood decoding," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2661–2675, Jun. 2007.

[79] B. Shim and I. Kang, "Sphere decoding with a probabilistic tree pruning," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4867–4878, Oct. 2008.

[80] T. Cui, S. Han, and C. Tellambura, "Probability-distribution-based node pruning for sphere decoding," *IEEE Trans. Veh. Technol.*, vol. 62, no. 4, pp. 1586–1596, May 2013.



RAZVAN-ANDREI STOICA (S'19) received the B.Sc. degree in electrical and computer engineering and the M.Sc. degree in communications, systems and electronics from Jacobs University Bremen, Germany, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree in electrical engineering with a focus on wireless communications systems. His research interests include the areas of communications, signal processing, and machine learning, such as communications systems, estimation theory, information theory, statistical processing, statistical learning, and adaptive filtering.

GIUSEPPE THADEU FREITAS DE ABREU (S'99–M'04–SM'09) received the B.Eng. degree in electrical engineering and the Specialization (*Latu Sensu*) degree in telecommunications engineering from the Universidade Federal da Bahia (UFBa), Salvador, Brazil, in 1996 and 1997, respectively, and the M.Eng. and Ph.D. degrees in physics, electrical and computer engineering from Yokohama National University, Japan, in 2001 and 2004, respectively. He was a Postdoctoral Fellow and then an Adjunct Professor (Docent) of statistical signal processing and communications theory with the Department of Electrical and Information Engineering, University of Oulu, Finland, from 2004 to 2006 and from 2006 to 2011, respectively. Since 2011, he has been a Professor of electrical engineering with Jacobs University Bremen, Germany. Since 2015, he has been, simultaneously, a Professor of electrical engineering with Ritsumeikan University, Japan. His research interests include a wide range within communications and signal processing, including communications theory, estimation theory, statistical modeling, wireless localization, cognitive radio, wireless security, MIMO systems, energy harvesting networks, random networks, and connected vehicles networks. His coauthored articles short-listed for best paper awards at the Asilomar Conference on Signals Systems and Computers, in 2009, 2012, and 2014. In 2000, he received the Uenohara Award from Tokyo University for his graduate work. He received the prestigious JSPS, Heiwa Nakajima, and NICT Fellowships, in 2010, 2013, and 2015, respectively. He was a co-recipient of the Best Paper Award at the Workshop on Positioning Navigation and Communications, in 2012 and 2014. He has served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, from 2009 to 2014. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS.

TAKANORI HARA (S'17) received the B.Eng. and M.Eng. degrees in engineering from The University of Electro-Communications, Tokyo, Japan, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree. His current research interests include communication theory, channel coding, and information theory.

KOJI ISHIBASHI (S'01–M'07) received the B.E. and M.E. degrees in engineering from The University of Electro-Communications, Tokyo, Japan, in 2002 and 2004, respectively, and the Ph.D. degree in engineering from Yokohama National University, Yokohama, Japan, in 2007. From 2007 to 2012, he was an Assistant Professor with the Department of Electrical and Electronic Engineering, Shizuoka University, Hamamatsu, Japan. From 2010 to 2012, he was a Visiting Scholar with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. Since 2012, he has been with the Advanced Wireless and Communication Research Center (AWCC), The University of Electro-Communications, where he is currently an Associate Professor. His current research interests include energy harvesting communications, wireless power transfer, channel codes, signal processing, and information theory.