G-MultiSphere: Generalizing Massively Parallel Detection for Non-Orthogonal Signal Transmissions

Chathura Jayawardena, *Student Member, IEEE*, and Konstantinos Nikitopoulos, *Member, IEEE*

***Abstract*— The increasing demand for connectivity and throughput, despite the spectrum limitations, has triggered a paradigm shift towards non-orthogonal signal transmissions. However, the complexity requirements of near-optimal detection methods for such systems becomes impractical, due to the large number of mutually interfering streams and to the rank-deficient or ill-determined nature of the corresponding interference matrix. This work introduces g-MultiSphere; a generic massively parallel and near-optimal sphere-decoding-based approach that, in contrast to prior work, applies to both well- and ill-determined non-orthogonal systems. We show that g-MultiSphere is the first approach that can support large uplink multi-user MIMO systems with numbers of concurrently transmitting users that exceed the number of receive antennas by a factor of two or more, while attaining throughput gains of up to 60% and with reduced complexity requirements in comparison to known approaches. By eliminating the need for sparse signal transmissions for non-orthogonal multiple access (NOMA) schemes, g-MultiSphere can support more users than existing systems with better detection performance and practical complexity requirements. In comparison to state-of-the-art detectors for NOMA schemes and non-orthogonal signal waveforms (e.g., SEFDM) g-MultiSphere can be up to an order of magnitude less complex, and can provide throughput gains of up to 60%.**

***Index Terms*— Sphere decoding, non-orthogonal-multiple-access (NOMA), multiple-input-multiple-output (MIMO), parallel processing.**

I. INTRODUCTION

THE next generations of communication systems are expected to provide enhanced throughput and massive connectivity with low latency requirements. These requirements have introduced a paradigm shift towards non-orthogonal transmission schemes. In this context, multi-user (MU) MIMO systems with aggressive spatial multiplexing and multicarrier, code-domain non-orthogonal-multiple-access (NOMA) schemes such as low-density-signature-OFDM (LDS-OFDM) and sparse-code-multiple-access (SCMA) have been of recent research interest [2], [3]. In another context, non-orthogonal faster-than-Nyquist [4] and spectrally-efficient-FDM (SEFDM) sacrifice signal orthogonality to achieve increased throughput and spectral efficiency [5]–[7].

However, to deliver these theoretical gains in practice, efficient schemes to demultiplex a large number of mutually interfering streams are necessary. In addition to the high dimensionality of such a detection problem, the interference matrix of recently proposed non-orthogonal transmission schemes is either ill-determined (e.g., SEFDM) or even rank-deficient (e.g., LDS-OFDM, power domain NOMA). To demultiplex the corresponding information streams, recently proposed NOMA solutions such as LDS-OFDM and SCMA employ sparse signal transmissions, that enables efficient detection by means of the Message Passing Algorithm (MPA) [8]. Still, the computation complexity of the corresponding messages (per iteration) is determined by the number of mutually interfering streams and the modulation order. In addition, and as we show in Section V, a high number of iterations are necessary to obtain accurate soft information for high order modulation schemes or codebooks. Furthermore, MPA does not apply to non-sparse structures such as simple power domain-NOMA [9], MIMO or SEFDM.

For non-sparse signals Sphere Decoding (SD) and its soft-output versions have been introduced as methods to reduce the complexity of Max-Log MAP detection [10], [11]. However, the latency requirements for obtaining exact Max-Log MAP soft information using depth first SDs [10] are random and become impractical even for full rank high dimensional systems. In a similar manner, the complexity of existing approximate fixed latency SD schemes, such as the Soft Fixed Complexity SD (SFSD) [12] and the K-best list SD [13], do not scale efficiently for large rank-deficient systems and their processing complexity becomes impractical. This is because, in principle, such approaches do not account for the specific interference matrix realization, but target the worst case transmission condition. As a result, list based approaches such as the K-best SD require large K values, and extensive long sorting operations that compromise their implementation efficiency.

In SEFDM systems, approximate SD based detection schemes have been adopted together with tailored preprocessing schemes that mild the complexity increase introduced by the corresponding ill-conditioned interference matrix.

Manuscript received April 3, 2019; revised August 28, 2019; accepted October 6, 2019. Date of publication October 28, 2019; date of current version February 14, 2020. This work was supported by the UK's Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/M029441/1. This article was presented in part in [1]. The associate editor coordinating the review of this article and approving it for publication was O. Oyman. *(Corresponding author: Konstantinos Nikitopoulos.)*

The authors are with the 5G Innovation Centre, Institute for Communication Systems (ICS), University of Surrey, Guildford GU2 7XH, U.K. (e-mail: c.jayawardena@surrey.ac.uk; k.nikitopoulos@surrey.ac.uk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2019.2949812

In this direction, the Truncated Singular Value Decomposition (TSVD) [14] has been applied as a preprocessing stage for approximate SDs. This approach, however, sacrifices optimality and results in an irreversible performance loss. The hard SD method together with the iterative preprocessing in [15] does not sacrifice optimality, but results in impractical complexity requirements for higher bandwidth compression, dense constellations and/or fading channels. To further improve the achievable throughput the authors of [16] introduced an iterative soft detection approach that is of very low complexity but, as we show in Section V, performs poorly for dense constellations.

An ideal detection scheme should be generic and applicable to any kind of non-orthogonal system employing both sparse and non-sparse signals [17], while efficiently mitigating the deficient or ill-determined rank nature of the interference matrix. In addition, such a detection scheme must be of very low latency and complexity even for a large number of non-orthogonal streams, to cope with the requirements of state-of-the-art systems [18]. Still, the processing requirements of existing detection approaches can easily exceed the capabilities of traditional processors [19], preventing the practical realization of large non-orthogonal systems.

The recently proposed MultiSphere SD framework enables practical and low latency massively parallel processing for large MIMO systems [20], [21]. MultiSphere focuses the available processing power on the most "promising" vector solutions. To achieve this, a "Metric-of-Promise"(MoP) is introduced that exploits the MIMO interference matrix to identify the Relative Position Vectors (RPVs) of the most promising solutions prior to detection. These RPVs are identified by an approximate SD tree search that can be realized by means of a K-best approach with K being the number of available Processing Elements (PEs) as suggested in [20]. Then, these RPVs are de-mapped to symbols based on their Euclidean distance to the received vector.

While MultiSphere can efficiently parallelize the large MIMO detection problem, it is not applicable to rank-deficient, non-orthogonal systems. G-MultiSphere extends this framework to be applicable to any practical non-orthogonal signal transmission, even if rank-deficient. To enable sphere decoding in both rank-deficient and ill-determined rank systems, g-MultiSphere's preprocessing stage includes a QR decomposition of a regularized interference matrix. However, due to this regularization, the probability distribution of the equalized received observable is no longer Gaussian, which adversely affects the search for the most promising RPVs. To resolve this, in Section IV, we introduce a new MoP based on the actual probability distribution of the equalized received observable after regularization, which redefines the search for the most promising RPVs. Due to the high implementation complexity of this MoP we also introduce an approximate MoP which favours implementation. We evaluate the validity of this approximation both analytically and by simulations.

For a large number of mutually interfering streams, and for soft detection a larger number of candidate solutions needs to be examined than in the case of hard detection. This is because the corresponding soft information calculation consists of multiple constrained hard detection problems [10]. Due to this, MultiSphere's preprocessing, that identifies the RPVs, and has been proposed to be based on the K-best SD, becomes of high complexity due to the required large K values and the corresponding sorting operations. To alleviate this problem, g-MultiSphere introduces a new preprocessing approach that resolves these bottlenecks. In particular g-MultiSphere adjusts the K value based on the specific interference matrix realization. In addition, it avoids the extensive sorting operations, substantially reducing the preprocessing computational complexity. Furthermore, in Section IV-D we provide a link between the complexity required to process a number of RPVs and the achievable vector-error-rate performance when processing these RPVs. We also discuss how the complexity requirements of the proposed approach scale with the number of mutually interfering information streams for a target vector-error-rate degradation compared to the uncoded ML vector-error-rate. At the detection stage, g-MultiSphere employs an efficient RPV to symbol de-mapping procedure similarly to [20], which inherits it's favourable complexity efficiency. As a result, to the best of our knowledge, g-MultiSphere is the first massively parallel detection approach that is scalable to rank-deficient large MIMO systems while attaining a processing latency similar to that of highly-suboptimal linear detectors. Thus, g-MultiSphere enables a generic computational framework that can be used on any non-orthogonal transmission scheme. We further emphasize that the proposed work fills a gap in joint detection of a large number of mutually interfering information streams, in the case of rank-deficient or ill-determined interference matrices. In particular, due to its near-optimal detection performance, the proposed approach can realize the potential of non-orthogonal waveforms such as SEFDM, currently left unexploited.

In Section V we show that even for conventional, 16×16, 16-QAM spatially multiplexed MU-MIMO systems, g-MultiSphere provides complexity gains of up to 50% compared to the originally proposed MultiSphere, without compromising the provided detection performance since it can better cope with ill-conditioned channel realizations. We also show that g-MultiSphere can efficiently support more users than twice the number of receiver antennas in a large multi-user MIMO environment while providing throughput gains of up to 60% in comparison to known approaches. This is achieved by just exploiting the inherent ability of the MIMO channel to support multiple users, and without applying any specific NOMA approach or specifically optimized multi-user codewords. In addition, g-MultiSphere can also achieve throughput gains of up to 60% in comparison to state-of-the-art soft detectors for SEFDM transmissions exploring unexploited capacity gains [6]. In Section V we also show that, when g-MultiSphere is applied to the detection of sparse LDS-OFDM signals, it can reduce both complexity and latency by more than an order of magnitude compared to MPA while providing improved throughput. By eliminating the need for sparse signal transmissions as specifically designed for MPA receiver processing, we show that g-MultiSphere can support many users per resource element, enabling overloading factors beyond two with practical complexity requirements.

In addition, we show that g-MultiSphere's preprocessing approach can reduce the corresponding complexity by up to an order of magnitude, for all the examined non-orthogonal schemes.

The rest of the paper is structured as follows. In Section II we introduce a generic model to describe non-orthogonal systems. In Section III we provide a primer on Sphere-Decoding-based detection for non-orthogonal systems, and in Section IV we describe g-MultiSphere; the proposed massively parallel detection scheme, together with the new MoP and improved preprocessing method to identify the most promising vector solutions that substantially reduces the preprocessing overhead when a large number of processing elements are utilized. Finally, in Section V we evaluate g-MultiSphere in comparison to state-of-the-art detection methods when applied to MU-MIMO, SEFDM and LDS-OFDM systems.

II. Generic Non-Orthogonal Transmission Modeling

The baseband received signal for a non-orthogonal system can be given by

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{w}, \tag{1}$$

where $\mathbf{y}$ is the $N \times 1$ received vector, $\mathbf{s}$ is the $M \times 1$ transmitted symbol vector with elements belonging to a constellation $\mathcal{O}$, $\mathbf{w}$ is the $N \times 1$ additive white Gaussian noise vector with variance σ^2 and $\mathbf{H}$ is the interference matrix that differs per non-orthogonal system. In the rest of this Section we consider non-orthogonal systems with both single and multiple-antenna receivers.

Uplink spatially multiplexed MIMO systems: In an uplink spatially multiplexed MIMO system with M transmit antennas and N receive antennas, the corresponding $N \times M$ MIMO channel matrix is $\mathbf{H}_{\text{MIMO}}$ and the $N \times 1$ received signal vector $\mathbf{y}$ could be modelled as in (1).

SEFDM systems: An SEFDM [7] block consists of M complex symbols transmitted within a time period T. Each of these M complex symbols modulate a Non-Orthogonal subcarrier. The bandwidth compression factor α is defined as $\alpha = \Delta f T$, with Δf being the frequency spacing between subcarriers, and with $\alpha = 1$ corresponding to an orthogonal system (e.g., OFDM). Then, the $N \times 1$ received vector, consisting of the received signal at each non-orthogonal subcarrier, is given by

$$\mathbf{y} = \mathbf{B}^H \mathbf{C}_h \mathbf{F}_\alpha \mathbf{s} + \mathbf{w}, \tag{2}$$

where the $M \times M$ fractional IFFT matrix $\mathbf{F}_\alpha$ consists of the entries $F_\alpha[k,n] = \exp(j2\pi\alpha(k-1)(n-1)/M)/\sqrt{M}$ for $n, k = 1, \ldots, M$. The $M \times M$ matrix $\mathbf{C}_h$ is circulant with its first column being $\begin{bmatrix} h_0 \; h_1 \; \ldots \; h_{L-1} \; 0 \; \ldots \; 0 \end{bmatrix}^T$ and L being equal to the number of channel taps in the time domain. Matrix $\mathbf{B}$ represents an orthonormal base which spans the SEFDM signal space and could be computed using a Gram-Schmidt orthonormalisation procedure as in [22]. Consequently, the SEFDM interference matrix is $\mathbf{H}_{\text{SEFDM}} = \mathbf{B}^H \mathbf{C}_h \mathbf{F}_\alpha$, which is ill-determined as discussed in [14].

LDS-OFDM NOMA systems: In an LDS-OFDM system an (orthogonal) subcarrier is loaded with the signals of multiple users which are superimposed. The $\tilde{N} \times 1$ received signal vector for an LDS-OFDM system where $\tilde{N}$ orthogonal subcarriers are occupied by M users is given by

$$\mathbf{y} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \ldots & \mathbf{h}_M \end{bmatrix} \circ \begin{bmatrix} \mathbf{g}_1 & \mathbf{g}_2 & \ldots & \mathbf{g}_M \end{bmatrix} \mathbf{s} + \mathbf{w} \tag{3}$$

where $\mathbf{h}_l$ is the frequency domain channel for user l, $l \in [1, M]$ and $\mathbf{g}_l$ is the "sparse signature vector" for the user l, which consists of complex entries that define how the signal is spread over subcarriers [2]. These sparse signature vectors are selected by predefined codebooks as discussed in [23]. Therefore, in relation to (1) the interference matrix is $\mathbf{H}_{\text{LDS-OFDM}} = \begin{bmatrix} \mathbf{h}_1 \; \mathbf{h}_2 \; \ldots \; \mathbf{h}_M \end{bmatrix} \circ \begin{bmatrix} \mathbf{g}_1 \; \mathbf{g}_2 \; \ldots \; \mathbf{g}_M \end{bmatrix}$, and it is rank-deficient since $\tilde{N} < M$.

For a receiver equipped with R_x antennas the received signal vector becomes of length $R_x \tilde{N}$ and is given by

$$\mathbf{y} = \begin{bmatrix} \mathbf{h}_{1,1} & \mathbf{h}_{2,1} & \ldots & \mathbf{h}_{M,1} \\ \mathbf{h}_{1,2} & \mathbf{h}_{2,2} & \ldots & \mathbf{h}_{M,2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{1,R_x} & \mathbf{h}_{2,R_x} & \ldots & \mathbf{h}_{M,R_x} \end{bmatrix} \circ \begin{bmatrix} \mathbf{g}_1 & \mathbf{g}_2 & \ldots & \mathbf{g}_M \\ \mathbf{g}_1 & \mathbf{g}_2 & \ldots & \mathbf{g}_M \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_1 & \mathbf{g}_2 & \ldots & \mathbf{g}_M \end{bmatrix} \times \mathbf{s} + \mathbf{w}, \tag{4}$$

where $\mathbf{h}_{k,R_x}$ is the frequency domain channel for user k for the R_x^{th} receive antenna. Therefore the $R_x N \times M$ interference matrix could be expressed as

$$\mathbf{H}_{\text{LDS-OFDM}} = \begin{bmatrix} \mathbf{h}_{1,1} & \mathbf{h}_{2,1} & \ldots & \mathbf{h}_{M,1} \\ \mathbf{h}_{1,2} & \mathbf{h}_{2,2} & \ldots & \mathbf{h}_{M,2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{1,R_x} & \mathbf{h}_{2,R_x} & \ldots & \mathbf{h}_{M,R_x} \end{bmatrix} \circ \begin{bmatrix} \mathbf{g}_1 & \mathbf{g}_2 & \ldots & \mathbf{g}_M \\ \mathbf{g}_1 & \mathbf{g}_2 & \ldots & \mathbf{g}_M \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_1 & \mathbf{g}_2 & \ldots & \mathbf{g}_M \end{bmatrix}, \tag{5}$$

and it is again rank-deficient ($R_x \tilde{N} < M$) for the challenging scenarios we consider in Section V. We note that in relation to the general model in (1) $N = R_x \tilde{N}$.

III. Sphere Decoding for Non-Orthogonal Systems

As discussed in Section II, the interference matrix $\mathbf{H}$ could be ill-determined or rank-deficient. Tikhonov regularization [24] is a proven method for mitigating the effects of small eigenvalues of an ill-determined rank matrix and has also been applied to rank-deficient systems [25]. In particular, instead of performing a QR decomposition on $\mathbf{H}$ we employ the QR decomposition of [26] on the Tikhonov regularised matrix $\bar{\mathbf{H}}$

$$\bar{\mathbf{H}} \triangleq \begin{bmatrix} \mathbf{H} \\ \lambda \mathbf{I}_M \end{bmatrix} = \bar{\mathbf{Q}} \bar{\mathbf{R}} = \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix} \bar{\mathbf{R}}, \tag{6}$$

with the regularisation parameter $\lambda = \sigma / \mathrm{E}|s_l|$. Where $\bar{\mathbf{Q}}$ is a $(M+N) \times M$ orthonormal matrix with elements $\bar{Q}_{i,l}$, ($i \in [1, M+N]$ and $l \in [1, M]$) and $\bar{\mathbf{R}}$ is a $M \times M$ upper triangular matrix.

Then, the "hard" ML estimation problem can be expressed as

$$\hat{\mathbf{s}}_{ML} = \arg \min_{\mathbf{s} \in \mathcal{O}^M} \{ \|\tilde{\mathbf{y}} - \bar{\mathbf{R}}\mathbf{s}\|^2 - \lambda^2 \|\mathbf{s}\|^2 \}, \tag{7}$$

where $\tilde{\mathbf{y}} = \mathbf{Q}_1^H \mathbf{y}$ is an $M \times 1$ vector.

Since $\bar{\mathbf{R}}$ is an upper triangular matrix, finding the ML solution can be translated into a tree search of height M and branching factor $|\mathcal{O}|$. Each node at level l can be identified by its partial symbol vector $\mathbf{s}_l = [s_l, s_{l+1}, \ldots, s_M]$ which also determines, the path from the root to that node, as well as from its partial Euclidean distance (PD) which can be calculated recursively as $d(\mathbf{s}_l) = d(\mathbf{s}_{l+1}) + e(s_l)$ where $e(s_l)$ is the non-negative cost assigned to each branch,

$$e(s_l) = \left(\left| \tilde{y}_l - \sum_{k=l}^{M} \bar{R}_{lk} s_l \right|^2 + \lambda^2 (E_{s_{max}} - |s_l|^2) \right). \quad (8)$$

Here, $E_{s_{max}} = \max(|s_l|^2)$ is the maximum energy of symbols chosen from the constellation $\mathcal{O}$. Then the ML detection problem is equivalent to finding the vector $\mathbf{s}$ with minimum $d(\mathbf{s}_1)$. According to the Schnorr-Euchner (SE) [27] enumeration the nodes are visited in an ascending order of their $e(s_l)$. For non constant amplitude transmit symbol constellations the minimization problem in (7) differs from the traditional SD tree search in [11], [21], [28] due to the $\lambda^2(E_{s_{max}} - |s_l|^2)$ term. As a result, applying the SE enumeration without exhaustively calculating the PDs of all constellation symbols, and instead by using simple geometrical properties as in [11], is not anymore feasible. To cope with this problem, enumeration schemes similar to [29], [30] could be used. However, these techniques are highly sequential and unsuitable for parallel processing. In the rest of the paper we will discuss how we can efficiently cope with this issue in massively parallel detection approaches.

In practical systems that employ soft channel decoding approaches like LDPC, soft information is required in the form of Log Likelihood Ratios (LLRs). The LLR for the jth coded bit b_j is defined as in [10]

$$L(b_j) \triangleq \ln \left(\frac{P[b_j = +1|\mathbf{y}, \mathbf{H}]}{P[b_j = -1|\mathbf{y}, \mathbf{H}]} \right). \quad (9)$$

The computation of LLRs, when the Max-Log approximation is employed, involves multiple constrained ML searches [10]. In particular, the LLR for the jth coded bit b_j could be expressed as

$$\begin{aligned} L(b_j) \approx & \min_{\mathbf{s} \in S_j^{-1}} \left\{ \frac{1}{\sigma^2} \|\tilde{\mathbf{y}} - \bar{\mathbf{R}}\mathbf{s}\|^2 + \frac{\lambda^2}{\sigma^2}(M E_{s_{max}} - \|\mathbf{s}\|^2) \right\} \\ & - \min_{\mathbf{s} \in S_j^{+1}} \left\{ \frac{1}{\sigma^2} \|\tilde{\mathbf{y}} - \bar{\mathbf{R}}\mathbf{s}\|^2 + \frac{\lambda^2}{\sigma^2}(M E_{s_{max}} - \|\mathbf{s}\|^2) \right\} \\ = & \operatorname{sign}(x_j)(D_j^{\overline{ML}} - D^{ML}), \end{aligned} \quad (10)$$

where x_j is the jth entry of the ML solution's bit label and S_j^{-1}, S_j^{+1} are the subsets of possible symbol vectors with jth bipolar bit set to $-1, +1$ respectively. Here D^{ML} is the metric of the ML solution and $D_j^{\overline{ML}}$ is the minimum metric from subset $S_j^{\bar{x_j}}$ for bit j. For a detection approach such as the proposed, which generates a list of candidate solutions, the LLRs can be calculated according to Eq. (10) based on the distance metrics in the list of candidate solutions. If a $D_j^{\overline{ML}}$ for a particular bit is not found in the list of candidates, a clipped value can be used instead according to [31].

IV. G-MultiSphere's Design

Originally, the MultiSphere [20] framework targeted the ML problem in full rank MIMO systems. MultiSphere focuses the available processing power on the most "promising" SD tree paths to constitute the transmitted symbol vector. This is achieved by a (prior to the detection stage) preprocessing stage that, based on the specific channel realization, identifies the most promising tree paths to include the transmitted symbol vector. The likelihood of each tree path to constitute the transmitted symbol vector is characterized by a *metric of promise* (MoP) $\mathcal{M}$ that is a function of the specific channel realization and not of the received symbol. Still, the originally proposed MoP cannot be applied to non-orthogonal systems since it does not account for the effect of interference matrix regularization that is required in ill-determined non-orthogonal systems. By using this MoP, MultiSphere identifies the most promising tree paths by an approximate K-best SD tree search with $K = N_{PE}$ being the maximum number of examined candidate solutions as suggested in [20]. However, a much larger number of tree paths (e.g., candidate solutions) needs to be processed in parallel in "soft" detection systems than in "hard" detection systems, due to the corresponding multiple constrained ML searches (see Section III). Then, the originally proposed K-best-based preprocessing (with N_{PE} being equal to the number of examined tree paths) can become impractical since it requires long sorting operations of the order of $O(N_{PE}|\mathcal{O}|\log\{N_{PE}|\mathcal{O}|\})$ and metric calculations of the order of $O(N_{PE}|\mathcal{O}|)$ per SD level. The preprocessing stage of MultiSphere finds the most promising tree paths by means of ordered distances to the received observable. Then, during the detection stage, when the actual signal is received, these tree paths need to be de-mapped to actual symbol vectors, in order to calculate the LLRs by employing Eq. (10). For this, MultiSphere applies an efficient procedure to de-map those paths onto actual symbols which avoids exhaustively calculating the corresponding distances for all symbols and sorting them.

In this Section we show how MultiSphere can be extended to ill-determined non-orthogonal systems. In particular, in Section IV-A a new MoP to [20] is derived for generic non-orthogonal systems that utilize the regularized QR decomposition introduced in Section III. Then, in Section IV-B, we introduce a novel efficient preprocessing module which identifies the most promising tree paths ($\tilde{N}_{PE} \leq N_{PE}$), based on the specific interference matrix realization, in a computationally efficient manner while avoiding any sorting operations as in K-best approaches. Finally, in Section IV-C we show that MultiSphere's tree path to symbol de-mapping procedure [20] can still, be approximately used by g-MultiSphere, despite the modified metric introduced in (8).

A. G-MultiSphere's MoPs

Similar to [20], a tree path is described by means of its ordered (in terms of PDs) position of its nodes to the received observable by an $M \times 1$ *relative position vector* (RPV) $\mathbf{k}$, with integer elements k_l ($l \in [1, M]$ and $k_l \in [1, |O|]$). Then, for the corresponding tree path, the node at level l is the k_l^{th} closest node to the received observable $\tilde{y}_l$.

Due to the regularization of the interference matrix, the received vector $\tilde{\mathbf{y}}$ includes a residual self-interference term that together with the noise [26] they form an effective noise term

$$\bar{\mathbf{w}} = \mathbf{Q}_1^H \mathbf{w} - \lambda \mathbf{Q}_2^H \mathbf{s}. \tag{11}$$

The SD detection is initiated at level M and, therefore, the corresponding equalized observable is

$$\hat{y}_M = \tilde{y}_M / \bar{R}_{M,M} = s_M^t + \hat{w}_M, \tag{12}$$

where s_M^t is the transmitted symbol at level M, and $\hat{w}_M = \bar{w}_M / \bar{R}_{M,M}$.

We first consider the probability of the symbol with the p^{th} smallest PD, denoted as x_p, being the transmitted symbol ($P_M[x_p = s_M^t]$).

For this, we consider the probability of noise being within the decision boundaries of x_p.

$$\begin{aligned} &P_M[x_p = s_M^t] \\ &\quad \approx P[\sqrt{\tilde{d}_M(p)} \le |\hat{w}_M| < \sqrt{\tilde{d}_M(p+1)}] \\ &\quad = P[\sqrt{\tilde{d}_M(p)} \le |\hat{w}_M|] - P[\sqrt{\tilde{d}_M(p+1)} < |\hat{w}_M|], \end{aligned} \tag{13}$$

where the the p^{th} smallest PD is denoted as $\tilde{d}_M(p)$. Here $P_M[x_p = s_M^t]$ has been simplified by employing the CDF properties [32]. Then, the considered probability ($P_M[x_p = s_M^t]$) can be bounded by

$$\begin{aligned} P_M[x_p = s_M^t] &\le P[\sqrt{\tilde{d}_M(p)} \le |\hat{w}_M|] \\ &= \sum_{\mathbf{s}\in\mathcal{O}^M} P\Big[\sqrt{\tilde{d}_M(p)} \le |\hat{w}_M| \Big| \mathbf{s}\Big] P[\mathbf{s}]. \end{aligned} \tag{14}$$

Since $\hat{w}_M$ is a function of two random variables, we consider the joint probability distribution of $\mathbf{w}$ and $\mathbf{s}$. Then, $\hat{w}_M$ can be considered to be Gaussian distributed with mean of $\sum_{l=1}^M \bar{Q}_{M+l,M}^H \lambda s_l / \bar{R}_{M,M}$ that is a function of $\mathbf{s}$. Therefore, the distribution of probability $P\Big[\sqrt{\tilde{d}_M(p)} \le |\hat{w}_M| \Big| \mathbf{s}\Big]$ can be modelled as a Rician CDF with mean $\mu_M(\mathbf{s}) = |\sum_{l=1}^M \bar{Q}_{M+l,M}^H \lambda s_l / \bar{R}_{M,M}|$ and a variance of $\tilde{\sigma}_M^2 = \sigma^2(1 - \sum_{l=1}^M \bar{Q}_{M+l,M}^H \bar{Q}_{M+l,M}) / |\bar{R}_{M,M}|^2$. Therefore $P_M[x_p = s_M^t]$ can be expressed as

$$\begin{aligned} &\sum_{\mathbf{s}\in\mathcal{O}^M} P\Big[\sqrt{\tilde{d}_M(p)} \le |\hat{w}_M| \Big| \mathbf{s}\Big] P[\mathbf{s}] \\ &= \frac{1}{|\mathcal{O}|^M} \sum_{\mathbf{s}\in\mathcal{O}^M} e^{-\frac{(\mu_M(\mathbf{s})^2 + \tilde{d}_M(p))}{\tilde{\sigma}_M^2}} \sum_{q=0}^{\infty} \Big(\frac{\mu_M(\mathbf{s})}{\tilde{d}_M(p)}\Big)^q I_q\Big(\frac{\mu_M(\mathbf{s}) \tilde{d}_M(p)}{\tilde{\sigma}_M^2}\Big) \\ &= \frac{1}{|\mathcal{O}|^M} \sum_{\mathbf{s}\in\mathcal{O}^M} \mathcal{Q}_1\Big(\frac{\mu_M(\mathbf{s})}{\tilde{\sigma}_M}, \frac{\sqrt{\tilde{d}_M(p)}}{\tilde{\sigma}_M}\Big), \end{aligned} \tag{15}$$

where $\mathcal{Q}_1$ is a first order Marcum Q-function and I_q is a modified Bessel function of the first kind [32]. We then consider the probability of the symbol vector corresponding to RPV $\mathbf{k}$, which is $\mathbf{x_k}$ being the transmitted symbol vector ($P[\mathbf{x_k} = \mathbf{s}^t]$). This probability can be bound as

$$\begin{aligned} &P[\mathbf{x_k} = \mathbf{s}^t] \\ &\quad = \prod_{l=1}^M P\Big[x_{k_l} = s_l^t \Big| \bigcap_{q=l+1}^M x_{k_q} = s_q^t\Big] \le \prod_{l=1}^M P_l[x_{k_l} = s_l^t] \\ &\quad \le \prod_{l=1}^M \frac{1}{|\mathcal{O}|^M} \sum_{\mathbf{s}\in\mathcal{O}^M} \mathcal{Q}_1\Big(\frac{\mu_l(\mathbf{s})}{\tilde{\sigma}_l}, \frac{\sqrt{\tilde{d}_l(p)}}{\tilde{\sigma}_l}\Big) \end{aligned} \tag{16}$$

In [20] it has been shown that the p^{th} smallest squared distance can be approximated by a linear function ($\tilde{d}_l(p) = c(p-1)$), with c depending on the minimum distance between constellation points. In particular, c can be determined by a linear approximation of the sorted squared distance from an inner constellation point to other constellation points. We note that, for a constellation with a minimum symbol distance of two, c can be set to 1.1 as specified in [20]. Now we define the MoP ($\bar{\mathcal{M}}(\mathbf{k})$) based on the actual CDF in Eq. (15) for non-orthogonal systems in relation to this probability, which can be calculated recursively as

$$\bar{\mathcal{M}}(\mathbf{k}_l) = \bar{\mathcal{M}}(\mathbf{k}_{l+1}) - \ln\Big\{\frac{1}{|\mathcal{O}|^M} \sum_{\mathbf{s}\in\mathcal{O}^M} \mathcal{Q}_1\Big(\frac{\mu_l(\mathbf{s})}{\tilde{\sigma}_l}, \frac{\sqrt{c(k_l - 1)}}{\tilde{\sigma}_l}\Big)\Big\}. \tag{17}$$

This MoP can be of impractical complexity to compute for large M or $|\mathcal{O}|$ values due to the exhaustive computation of the Q-functions of the order $|\mathcal{O}|^M$ that are required to determine the actual CDF of $|\hat{w}_M|$ in (15). Therefore, in the rest of this paper, we introduce and employ an approximation that favours implementation, and as we show in Section V, results in a negligible performance loss. Note that $\hat{w}_M$ constitutes of the two independent random variables $\mathbf{w}$ and $\mathbf{s}$.

$$\hat{w}_M = \frac{\sum_{l=1}^M \bar{Q}_{l,M}^H w_l}{\bar{R}_{M,M}} + \frac{\sum_{l=1}^M \bar{Q}_{M+l,M}^H \lambda s_l^t}{\bar{R}_{M,M}} \tag{18}$$

The first term in the right hand side of (18) has a Gaussian distribution with zero mean and variance $\sigma^2(1 - \sum_{l=1}^M \bar{Q}_{M+l,M}^H \bar{Q}_{M+l,M}) / |\bar{R}_{M,M}|^2$. According to the central limit theorem, the second term tends to become Gaussian with zero mean and variance $\sigma^2(\sum_{l=1}^M \bar{Q}_{M+l,M}^H \bar{Q}_{M+l,M}) / |\bar{R}_{M,M}|^2$ for large M values. Then the CDF of $|\hat{w}_M|$ can be considered as a sum of Gaussian distributed random variables, and the probability $P[\sqrt{\tilde{d}_M(p)} \le |\hat{w}_M|]$ can be approximated by a Rayleigh CDF as

$$P[\sqrt{\tilde{d}_M(p)} \le |\hat{w}_M|] \approx e^{-\frac{\tilde{d}_M(p)|\bar{R}_{M,M}|^2}{\sigma^2}} \tag{19}$$

This approximation has been validated by simulating the CDF of $|\bar{w}_M|$ for the high dimensional systems considered in Section V and the SNRs of interest (See Fig. 5). Based on this approximate CDF, we define the simplified MoP ($\mathcal{M}(\mathbf{k})$) for non-orthogonal systems as

$$\mathcal{M}(\mathbf{k}) = \sum_{l=1}^M \frac{c(k_l - 1)|\bar{R}_{ll}|^2}{\sigma^2} \le -\ln\{P[\mathbf{x_k} = \mathbf{s}^t]\}. \tag{20}$$

Therefore the tree paths are visited according to this upper bound of their probabilistic likelihood. The MoPs could be calculated recursively by

$$\mathcal{M}(\mathbf{k}_l) = \mathcal{M}(\mathbf{k}_{l+1}) + \frac{c(k_l - 1)|\bar{R}_{l,l}|^2}{\sigma^2}, \tag{21}$$

where $\mathbf{k}_l$ is again the partial RPV and $\mathcal{M}(\mathbf{k}_l)$ is the partial MoP of $\mathbf{k}$ at level l. Assuming the term $c|\bar{R}_{l,l}|^2/\sigma^2$ is precomputed, this MoP calculation consists only of an addition and a real integer multiplication in contrast to the computationally intensive MoP in Eq. (17) based on the exact CDF. For the rest of this paper this MoP definition is applied unless otherwise denoted.

B. G-MultiSphere's Preprocessing Stage

The purpose of the preprocessing stage is to identify the most promising RPVs.

To reduce the search space of the preprocessing stage we consider a probabilistic threshold for the MoPs that determines the number of required RPVs $\tilde{N}_{PE}$. The number of required RPVs ($\tilde{N}_{PE} \leq N_{PE}$) depend on the specific interference matrix realization and correspond to the number of required PEs.

In particular, based on the bound of (20), we will ignore candidate solutions (i.e., prune nodes) with an upper bound of prior probability of being correct smaller than a predefined threshold (P_{th}). To achieve this we only consider the $\tilde{N}_{PE}$ RPVs that satisfy

$$\mathcal{M}(\mathbf{k}) \leq -\ln\{P_{th}\} = \mathcal{M}_{th}. \tag{22}$$

We note that P_{th} is a design parameter. As we discuss in Section IV-D P_{th} is related to the probability of including the correct solution in the examined $\tilde{N}_{PE}$ RPVs. As we further analyze in Section IV-D, P_{th} can determine how closely we can approach the ML vector-error-rate.

Since the recursive structure of the MoP calculation in (21) resembles that of an SD, the search for RPVs translates into a tree search. Therefore the RPVs with the smallest MoPs are identified by a K-Best SD-like tree search where each node at level l is characterized by a partial RPV $\mathbf{k}_l$ and a partial MoP $\mathcal{M}(\mathbf{k}_l)$ similar to the description in Section III. However, as we will explain later, our approach results in a different $\tilde{N}_{PE}$ value per tree level and does not require any sorting operations.

The selection of the most promising RPVs starts at the highest tree level. Then, the tree nodes are visited in the ascending order of their $\mathbf{k}_M$ indices and all the nodes with a partial MoP larger than $\mathcal{M}_{th}$ are pruned. For each of the survived nodes at level M, the child node with the smallest partial MoP is expanded first. At level $M-1$ the expanded nodes are visited in an ascending order of their parent's partial MoPs and all those with a partial MoP larger than $\mathcal{M}_{th}$ are pruned. The approach continues with expanding the child node with the second partial MoP for each of the survived nodes at level M. Those nodes with partial MoPs larger than $\mathcal{M}_{th}$ are also pruned. The node expansion continues until either all nodes at level $M-1$ are examined or until when the number of non-pruned nodes reaches N_{PE}. Then, the same procedure is applied for the rest of the tree levels. An example of the proposed MoP identification method is shown in Fig. 1. The process starts at level M.

Algorithm 1 Pseudocode for the Preprocessing Stage to Identify the RPVs

Inputs: $\bar{\mathbf{R}}$, M, $|\mathcal{O}|$
$l \leftarrow M$ where l denotes the current level
$\mathbf{P}$ is the $M \times \tilde{N}_{PE}$ RPV matrix which stores the $\tilde{N}_{PE}$ RPVs. The elements of $\mathbf{P}$ are integers taking values from 1 to $|\mathcal{O}|$
while $l > 1$ **do**
  for $n \leftarrow 1, |\mathcal{O}|$ **do**
    for $j \leftarrow 1, \tilde{N}_{PE}$ **do**
      Expand the n^{th} child node of j^{th} parent node in current $\mathbf{P}$ and compute its partial MoP.
      if Partial MoP $\leq \mathcal{M}_{th}$ **then**
        Update rows M to $l-1$ of new RPV matrix $\mathbf{P}$ with the selected RPV.
      end if
    end for
    if Number of selected RPVs $\geq N_{PE}$ **then**
      break
    end if
  end for
  $l \leftarrow l - 1$
end while
Output: $\mathbf{P}$

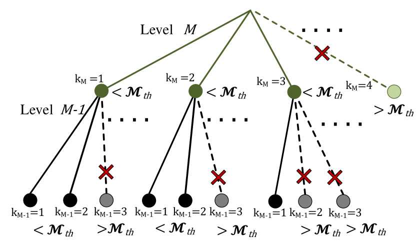

Fig. 1. Tree path selection example.

First, we compare the partial MoPs of all nodes with the probabilistic threshold $\mathcal{M}_{th}$. Where the fourth node ($\mathbf{k}_M = 4$) is pruned since we assume that the corresponding partial MoP exceeds the probabilistic threshold. Then, only three partial RPVs ($\mathbf{k}_M = 1, 2, 3$) remain as survivors for the next level (level $M-1$). At level $M-1$, the first child node for each survived node is expanded. Then, all expanded nodes are selected since their MoPs are within the threshold. Next, the second child node for each of the survived nodes at level M is expanded. In this case, the second child node of the third parent node (partial RPV $\mathbf{k}_{M-1} = [2, 3]^T$) is not selected since we assume that its partial metric exceeds $\mathcal{M}_{th}$. Subsequently, the third child node for each of the survived nodes at level M is expanded, and those with partial MoPs larger than $\mathcal{M}_{th}$ are again pruned. Since all other nodes at level $M-1$ have been pruned, the process will continue by expanding the first child node of these survived nodes (partial RPVs $\mathbf{k}_{M-1} = [1, 1]^T, [1, 2]^T, [2, 1]^T, [2, 2]^T, [1, 3]^T$).

C. G-MultiSphere's De-Mapping

The preprocessing stage identifies a list of the most promising paths (RPVs) by means of ordered distances to the received

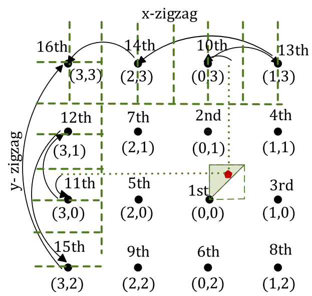

Fig. 2. Approximate pre-ordering for 16QAM. Here $\hat{y}_l$ is depicted by the pentagon.

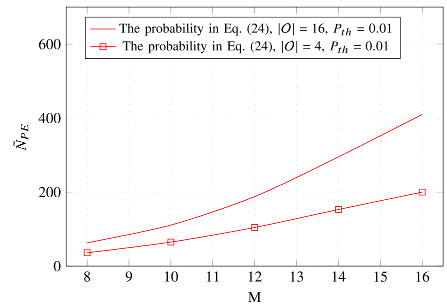

Fig. 3. The number of $\tilde{N}_{PE}$ RPVs required to be examined as in Eq. (24) for a considered P_{th} probability, in the case of $M \times M$ MIMO. An SNR of 12 dB is assumed for $|\mathcal{O}| = 16$ and an SNR of 7 dB is assumed for $|\mathcal{O}| = 4$.

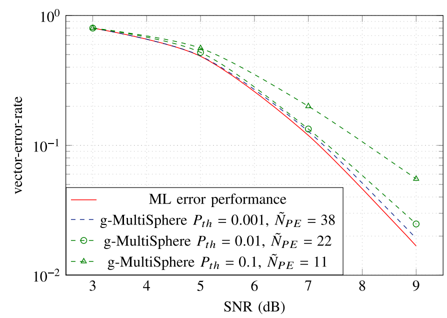

Fig. 4. The uncoded vector-error-rate performance of g-MultiSphere with several P_{th} values for a 4 QAM 8×8 MIMO system.

point $\hat{y}_l$, where

$$\hat{y}_l = (\tilde{y}_l - \sum_{j=l+1}^{n_t} \bar{R}_{l,j} s_j)\bar{R}_{ll}^{-1}. \tag{23}$$

During the detection stage we need to de-map these nodes onto actual symbols to calculate the LLRs by employing Eq. (10). This would traditionally require exhaustively calculating the corresponding distances for all symbols and sorting them, that would result in a substantial complexity overhead. Instead, MultiSphere [28] introduced a symbol mapping of two-dimensional zigzag coordinates. Then, an approximate symbol ordering relative to $\hat{y}_l$ (The pentagon in Fig. 2), could be predefined as a sequence of these two-dimensional zigzag coordinates [28]. Following the same principles we utilize the same preordering introduced in [28]. In Fig. 2 we describe the employed preordering. The zigzag coordinates in Fig. 2 could be identified based on the sectors containing the real and imaginary parts of $\hat{y}_l$. We note, that based on (8), the actual ordering should take place based on the corresponding partial distance $e(s_l)$ and not on the distance from the received symbol. This is because (8) also contains a term related to the energy of the constellation symbol (s_l) in addition to the distance from the received symbol. However, this is feasible since the second part in (8) has been modelled in as additional noise (see Eq. (11)) at the preprocessing stage.

D. Discussion on g-MultiSphere's Performance

In this Subsection we provide a link between the complexity required to process $\tilde{N}_{PE}$ RPVs and the provided vector-error-rate performance when processing $\tilde{N}_{PE}$ RPVs. In addition, we discuss how the complexity requirements of the proposed approach scale with the number of mutually interfering information streams for a target vector-error-rate degradation compared to the uncoded ML vector-error-rate.

Based on the description in Section IV-A, we can approximate the probability $P[\mathbf{s}^t \notin \mathcal{S}]$ of not having the transmitted symbol vector in the examined set of $\tilde{N}_{PE}$ RPVs by

$$P[\mathbf{s}^t \notin \mathcal{S}] = 1 - \sum_{\mathbf{k}=1}^{\tilde{N}_{PE}} P[\mathbf{x_k} = \mathbf{s}^t] \approx 1 - (1 - e^{-\mathcal{M}(\tilde{N}_{PE})}) = P_{th}, \tag{24}$$

where $\mathcal{M}(\tilde{N}_{PE})$ is the $\tilde{N}_{PE}^{th}$ largest MoP corresponding to the least probable solution in the examined set. According to the selection criteria for RPVs in Eq. (22) the probability $e^{-\mathcal{M}(\tilde{N}_{PE})}$ is equal to P_{th}. In Fig. 3, we show how the number of $\tilde{N}_{PE}$ RPVs required to be examined to approach a $P_{th} = 0.01$ probability scale with respect to M and $|\mathcal{O}|$. We note that the tree paths required to be examined by g-MultiSphere is much smaller than that theoretically required by the FSD. For an example, FSD requires the processing of 4096 tree paths for $|\mathcal{O}| = 16$ and for $M = 10, .., 16$ to achieve same diversity as ML detection [33]. These gains are also consistent with the coded comparisons in Section V.

In addition we can link the achievable error-rate by g-MultiSphere, from processing $\tilde{N}_{PE}$ RPVs, to ML error-rate. Therefore, we relate $P[\mathbf{s}^t \notin \mathcal{S}]$ to the error probability of g-MultiSphere. In particular the error probability can be expressed as

$$P[\hat{\mathbf{s}} \neq \mathbf{s^t}] = P[\hat{\mathbf{s}} \neq \mathbf{s}^t \cap \mathbf{s}^t \in \mathcal{S}] + P[\hat{\mathbf{s}} \neq \mathbf{s}^t \cap \mathbf{s}^t \notin \mathcal{S}], \tag{25}$$

where $\mathcal{S}$ is the set of tree paths or RPVs examined by g-MultiSphere and $\hat{\mathbf{s}}$ is its symbol estimate. The first part $P[\hat{\mathbf{s}} \neq \mathbf{s}^t \cap \mathbf{s}^t \in \mathcal{S}]$ corresponds to an ML error event. This can be simplified by considering the ML error probability as $P[\hat{\mathbf{s}} \neq \mathbf{s}^t \cap \mathbf{s}^t \in \mathcal{S}] \leq P[\hat{\mathbf{s}}_{ML} \neq \mathbf{s}^t]$. Therefore the error probability becomes

$$P[\hat{\mathbf{s}} \neq \mathbf{s^t}] \leq P[\hat{\mathbf{s}}_{ML} \neq \mathbf{s}^t] + P[\mathbf{s}^t \notin \mathcal{S}], \tag{26}$$

As we illustrate in Fig. 4, P_{th} determines how closely we can approach ML vector-error-rate. We note that $P_{th} = 0.01$ provides a good performance complexity trade-off and the corresponding performance degradation is negligible in coded

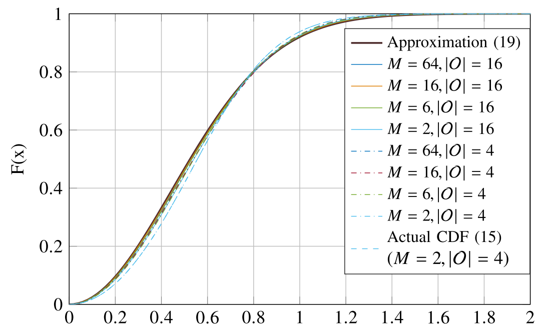

Fig. 5. The CDF of $|\bar{w}_M|$ in comparison to the Rayleigh approximation for various system dimensions at a 3dB SNR.

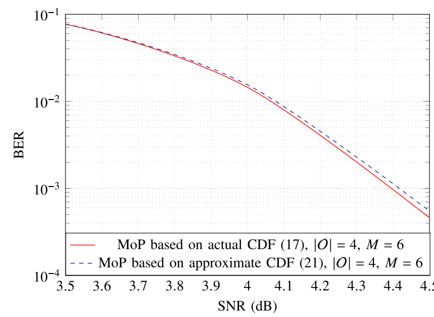

Fig. 6. BER performance of g-MultiSphere when employing the MoP based on the actual CDF (17) in comparison to the MoP based on the approximate CDF (21) for a 4 QAM 6×6 MIMO system.

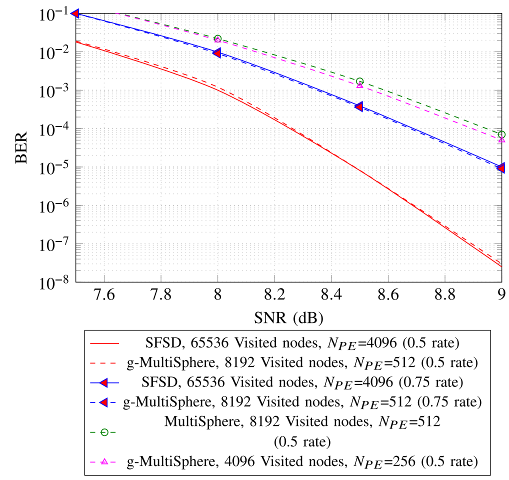

Fig. 7. Soft detection BER performance of g-MultiSphere in comparison with existing methods for a 16 QAM 16×16 MIMO system and several code rates.

systems due to the coding gain. Therefore $P_{th} = 0.01$ is considered in Section V.

In the previous paragraphs we have analyzed the achievable vector-error-rate when processing $\tilde{N}_{PE}$ PEs. Here we provide the processing complexity requirements per PE, and therefore the complexity required to achieve a given vector-error-rate. The number of multiplications required by g-MultiSphere at each PE is same as that of SFSD. In particular, g-MultiSphere requires $3\tilde{N}_{PE}M(1+(M+1)/2)$ real multiplications. The number of multiplications per PE has been obtained in a similar manner to [12]. We note that, there is also an additional implementation overhead for the RPV-to-symbol de-mapping (Section IV-C), that has been shown in [20] to be low. Since the detection stage of g-MultiSphere shares the same structure with the original MultiSphere, the implementation evaluation of [20] generally applies also to g-MultiSphere.

V. Simulation Evaluation

In this Section we first validate the approximations of the probability distributions discussed in Section IV-A. Then, we evaluate the performance of g-MultiSphere in three promising non-orthogonal transmission schemes. In particular, we consider spatially-multiplexed MIMO, SEFDM and LDS-OFDM (Section II). We compare both the complexity and the error-rate performance of the g-MultiSphere scheme with state-of-the-art detection methods for each non-orthogonal transmission scheme. In addition, to evaluate the efficiency of our proposed preprocessing stage (Section IV-B) we compare its complexity and provided performance with the original preprocessing approach introduced in [28].

A. Exact and Approximate MoP Evaluation

In this Section we validate the approximations of the probability distributions and evaluate the Bit Error Rate (BER) performance corresponding to the MoPs introduced in Section IV-A. As we verify in Fig. 5, the CDF of $|\bar{w}_M|$ (See Eq. (18)) approaches the Rayleigh CDF approximation in (19) when at least one of the parameters M or $|\mathcal{O}|$ is of high value. In Fig. 6 we evaluate the BER performance of g-MultiSphere with the MoPs introduced in Section IV-A, and in particular for the aforementioned actual and approximate probability distributions. As we discuss in Fig. 5, the CDF approximation results in a larger error for low M or $|\mathcal{O}|$ values. Here, we consider the lowest M and $|\mathcal{O}|$ values of our evaluated scenarios to account for the worst case effect of this approximation on the BER performance of g-MultiSphere. Fig. 6 shows the BER performance of g-MultiSphere when employing the MoP based on the actual CDF (17) in comparison to the MoP based on the approximate CDF (21) for a 4 QAM 6×6 MIMO system. We note that, the performance degradation due to the CDF approximation is negligible. Furthermore, as we discuss in Section IV-A, the exact MoP (17) based on the actual CDF requires an exhaustive computation of the Q-functions of the order $|\mathcal{O}|^M$. In contrast, the approximate MoP (21) based on the approximate CDF only requires an addition and a real integer multiplication. Therefore, we only focus on the approximate MoP (21) based on the approximate CDF in the rest of this paper.

B. Massively Parallel Regularized MIMO Detection

In this Section we evaluate and compare g-MultiSphere with state-of-the-art MIMO detectors, both in traditional ($N \geq M$) and overloaded ($N < M$) MIMO systems. In Fig. 7 we compare the BER performance and the complexity requirements in terms of visited nodes of g-MultiSphere with the soft-fixed-complexity-sphere-decoder (SFSD) [12], as well as with the

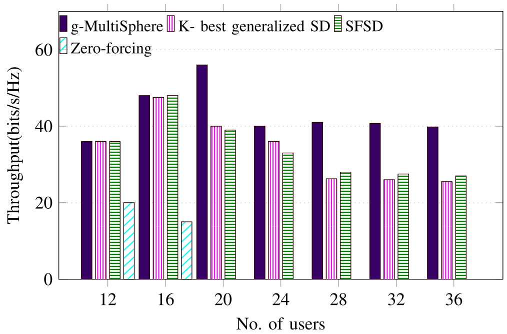

Fig. 8. The achievable throughput of g-MultiSphere, SFSD, soft K- best generalized SD and zero-forcing as the number of users transmitting to a 16-antenna base station increase at an 15 dB average SNR per user. The employed modulation order is chosen from 4, 16, 64 and code rate from 1/2, 2/3, 3/4, 5/6 to maximize throughput.

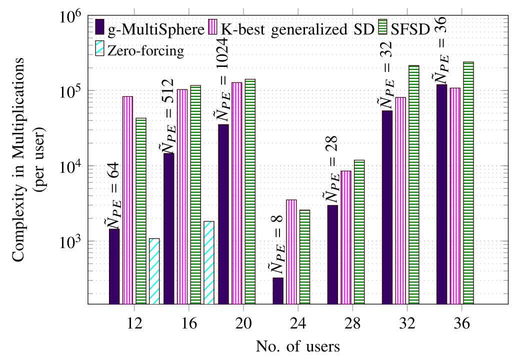

Fig. 9. The complexity requirements of g-MultiSphere, SFSD, soft K- best Generalized SD and zero-forcing to achieve the throughput depicted in Fig. 8.

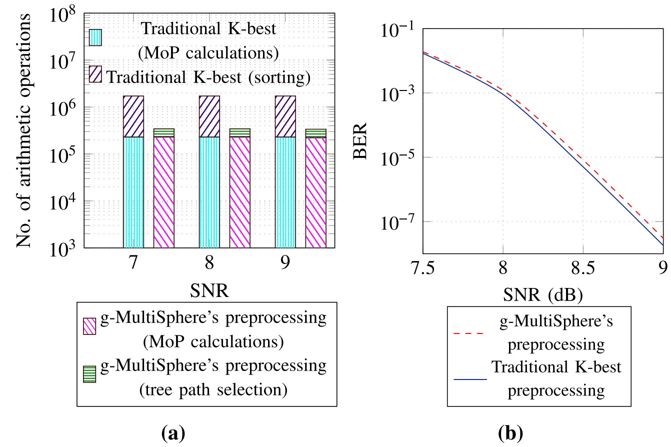

Fig. 10. a) The complexity of g-MultiSphere's preprocessing (Section IV-B) in comparison with MultiSphere's preprocessing ([28]), and b) the corresponding BER performance for the considered MIMO system with $N_{PE} = 512$.

approximate soft-extension of the recently proposed MultiSphere [28] that uses an unregularized, sorted QR decomposition. To the best of our knowledge these are the most efficient massively parallelizable detection approaches that could apply to large non-orthogonal systems. The processing complexity per visited node is similar for both the SFSD and g-MultiSphere as discussed in Section IV-D. Therefore, we use the number of visited nodes as a metric for our complexity comparisons. A 16 QAM modulated 16×16 MIMO-OFDM system is assumed with 52 active subcarriers where each subchannel between a transmit-receive antenna pair is modelled as a 4 tap i.i.d Rayleigh channel and 1944 block length LDPC codes are employed as in the 802.11 standard. As we show in Fig. 7, g-MultiSphere achieves a similar BER to SFSD with $1/8^{th}$ of the visited nodes, and these gains are consistent for different code rates. Fig. 7 also shows that despite the introduction of self-interference, the regularized sorted QR decomposition together with the proposed preprocessing results in a significant BER performance improvement. This is because the regularization can better cope with the ill-conditioned channel realizations.

In contrast to traditional MultiSphere, g-MultiSphere can also function in highly overloaded scenarios where the number of transmitting users is much higher than the available base station antennas. In Fig. 8 and Fig. 9, we investigate such scenarios. As we show in Fig. 8, g-MultiSphere can support number of users beyond twice the receiver antennas while providing two times the throughput of zero-forcing with 12 users. In addition, while supporting 36 users, g-MultiSphere can provide throughput gains of 60% in comparison to a K-best based generalized SD [25] when K is such that both approaches have similar complexity requirements. These complexity requirements are shown in Fig. 9 for all evaluated systems in terms of real multiplications per user. Still, it is significant to note that in contrast to the other approaches, K-best SDs would require additional computationally extensive sorting operations, that have not been considered in Fig. 9. We note that unlike g-MultiSphere, SFSD cannot focus the available processing power on the most promising tree paths, which results both in high complexity requirements and lower throughput.

In Fig. 10a we compare the preprocessing complexity of g-MultiSphere with the K-best based approach initially introduced in [28]. The tree path selection procedure of the proposed preprocessing requires only $O(\tilde{N}_{PE}|\mathcal{O}|)$ comparisons that can be executed in parallel, as discussed in Section IV-B in detail. This procedure replaces sorting and results in more than an order of magnitude complexity savings compared to a K-best approach for the N_{PE} values used in Fig. 10a. To account both for MoP calculations and the required sorting operations, here the comparisons are shown in terms of arithmetic operations. As we also show in Fig. 10b these complexity gains come with insignificant BER performance degradation.

C. Massively Parallel Regularized SEFDM Detection

Here we evaluate and compare g-MultiSphere with existing near-optimal SEFDM detectors. We consider both uncoded transmission with hard detection and coded transmission with soft detection.

In Fig. 11 we compare the uncoded hard detection BER of g-MultiSphere with K-best SDs for a 16 subcarrier system with $\alpha = 0.67$. The proposed method can approach ML error performance with $1/5^{th}$ of the complexity of state-of-the-art K-best SDs. Again, the additional complexity of the sorting operations required for the traditional K-best SDs has not

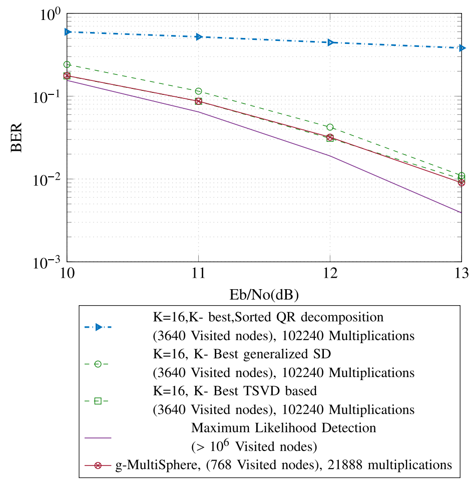

Fig. 11. Hard detection BER performance of g-MultiSphere in comparison with existing methods for a 16 QAM 16 subcarrier uncoded SEFDM system in AWGN channels and $\alpha = 0.67$. An $N_{PE} = 48$ maximum number of allocated PEs is assumed for g-MultiSphere.

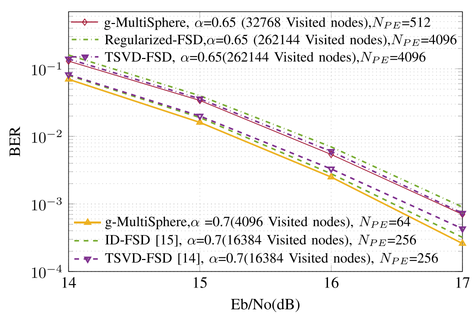

Fig. 12. Hard detection BER performance of g-MultiSphere in comparison with existing methods for a 16 QAM 64 subcarrier uncoded SEFDM system in AWGN channels.

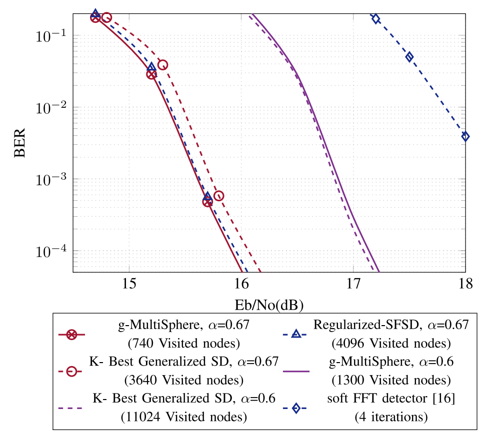

Fig. 13. Soft detection BER performance of g-MultiSphere in comparison with existing methods for a 16 QAM 16 subcarrier SEFDM system with $N_{PE} = 128$. Rayleigh fading channels and $1/2$ rate LDPC codes are assumed.

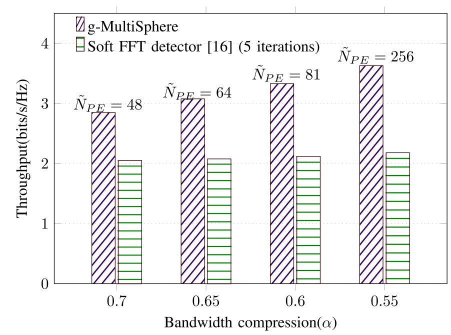

Fig. 14. The achievable throughput of g-MultiSphere in comparison to the state-of-the-art. Each method chooses the modulation scheme and code rate combination that maximizes throughput. An SNR of 17dB is assumed. An $N_{PE} = 256$ maximum number of allocated PEs is assumed for g-MultiSphere.

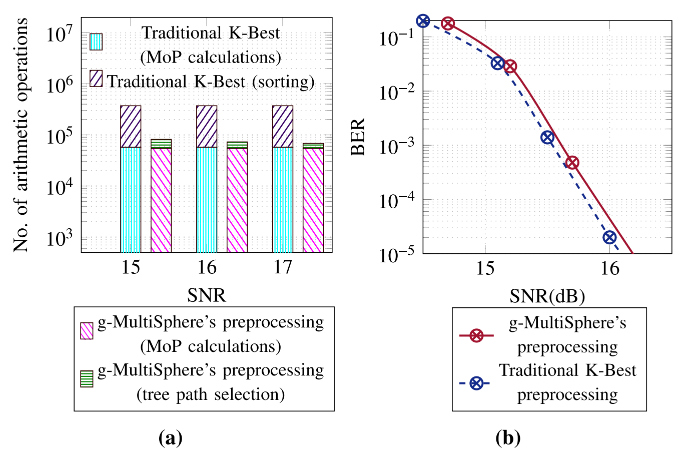

Fig. 15. a) Complexity of g-MultiSphere's preprocessing in comparison to MultiSphere's K-best approach, b) The corresponding BER performance, for 16 subcarrier SEFDM systems with $\alpha = 0.67$ and $N_{PE} = 128$.

been considered in the complexity comparisons of Fig. 11. In Fig. 12 we compare the BER of g-MultiSphere with TSVD-FSD [14], ID-FSD [15] and a GSD based Regularized FSD for 64 subcarrier systems. As shown in Fig. 12 g-MultiSphere can achieve a similar BER to existing FSDs with $1/8^{th}$ of the complexity when $\alpha = 0.65$.

Fig. 13 shows the soft detection complexity requirements (in terms of visited nodes) and the BER performance of g-MultiSphere compared to SFSD, soft K-best SD [12], [13] and the soft FFT detector of [16], for 16 subcarrier systems with $\alpha = 0.67$ and 0.6. It can be seen that g-MultiSphere's complexity gains increase when the overlap between subcarriers increases (when the value of parameter α decreases) and g-MultiSphere can be up to an order of magnitude less complex than the state-of-the-art when $\alpha = 0.6$.

To the best of our knowledge, g-MultiSphere is the first approach that can realize throughput trades consistent to what is expected from theory. This is shown in Fig. 14 where adaptive modulation and coding is assumed. It can also be shown that the Soft FFT detector of [16] benefits from a low detection complexity but it requires iterations with the channel decoder, resulting in a higher processing latency than g-MultiSphere. G-MultiSphere can achieve throughput gains of up to 60% in comparison to Soft FFT detection. Fig. 15a shows the complexity savings of g-MultiSphere's preprocessing in comparison to MultiSphere's K-best approach for the SEFDM system considered in Fig. 13. The proposed

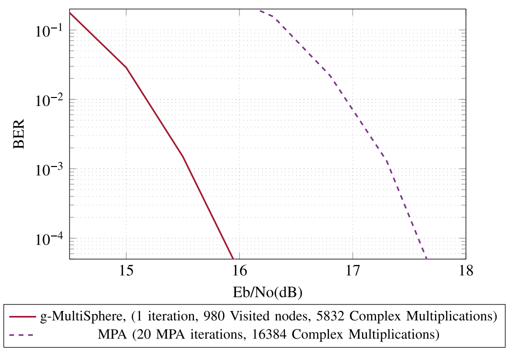

Fig. 16. Soft-Output BER performance of proposed method in comparison with Message Passing Algorithm (Rayleigh fading channels with $1/2$ rate LDPC coding). 6 users employing 16 point codebooks as in [23] and 4 subcarriers are assumed.

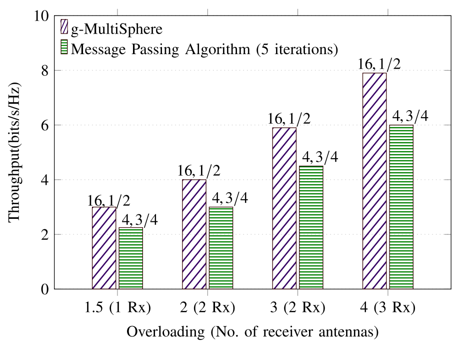

Fig. 17. The achievable throughput of g-MultiSphere in comparison to the state-of-the-art for LDS-OFDM in Rayleigh fading channels. Each method chooses the modulation scheme and code rate combination to maximize throughput. The number of assumed receiver antennas are indicated next to the overloading factor. Sparse codebooks are assumed at a SNR of 17dB. The codebook size and the code rate is indicated above the bar.

preprocessing reduces complexity by an order of magnitude while the corresponding performance loss is small.

D. Massively Parallel LDS-OFDM Detection

In Fig. 16 we compare g-MultiSphere with Max-Log MPA for a LDS-OFDM sytem where 6 users employ 4 subcarriers. The sparse signature matrix and the codebooks of [23] have been adopted. It can been seen that for this scenario, g-MultiSphere can provide an SNR gain of 1.5 dB compared to traditional MPA. For this signature matrix and interfering users the computational complexity of MPA is of $N|\mathcal{O}|^3$ per iteration, while g-MultiSphere requires only a PD calculation per visited node similarly to SFSD [12]. In particular, and as shown in Fig. 16, g-MultiSphere requires 5832 complex multiplications [12] while MPA requires 16384 complex multiplications per iteration, resulting in more than an order of magnitude reduction in overall complexity. In addition and in contrast to MPA, g-MultiSphere requires only one iteration and therefore it has a detection latency similar to that of the highly sub-optimal linear detection approaches.

G-MultiSphere can consistently exploit the throughput gains at highly overloaded scenarios as shown in Fig. 17 with gains of up to 35% in comparison to MPA. Here we consider the overloading factor as the ratio of the number of users to

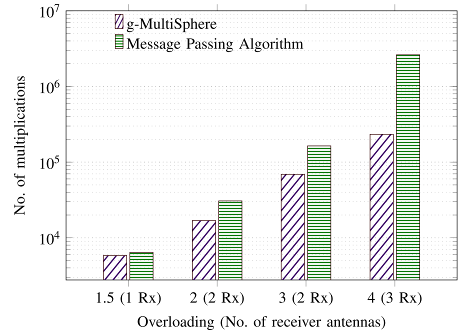

Fig. 18. The complexity requirements of g-MultiSphere in comparison to MPA with five iterations for the throughput results in Fig. 17. Here only the number of multiplications are considered as in [12] since they dominate the complexity.

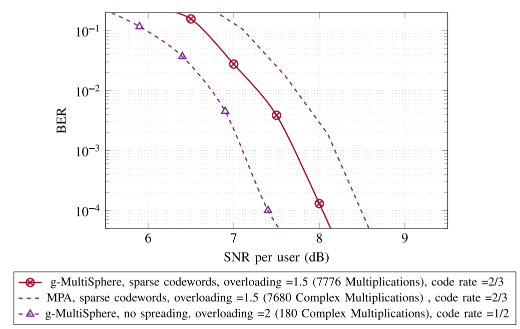

Fig. 19. Soft-Output BER performance of g-MultiSphere in comparison to existing methods for 4 point codebooks in Rayleigh fading channels. 6 users and 4 subcarriers are assumed for the overloading factor of 1.5 and 8 users are assumed for the overloading factor of 2 with each subcarrier simply occupied by 2 users. Sparse codebooks as in [23] are assumed.

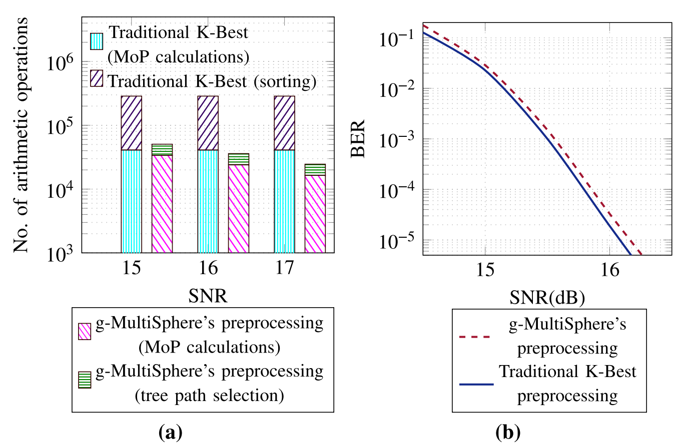

Fig. 20. a) Complexity of g-MultiSphere's preprocessing in comparison to MultiSphere's K-Best approach, b) the corresponding BER performance, for a 6 user LDS-OFDM systems employing 16 point codebooks.

the resource elements ($M/\tilde{N}$). In Fig. 18, we show that the complexity savings of g-MultiSphere are higher for higher overloading factors with reductions of more than an order of magnitude.

The proposed approach does not require sparse interference matrices ([23]) that are specifically designed for MPA receiver processing. In Fig. 19, we show that, for the same overall transmission rate, a g-MultiSphere based-scheme, with an overloading factor of 2 can support a larger number of users (8 instead of 6) and still provide an improved BER performance. As we show in Fig. 19, despite these gains, the detection complexity for this g-MultiSphere based-scheme is much lower than for sparse signal spreading due to the low dimensionality of the corresponding detection problem.

In Fig. 20a we compare the complexity savings of the g-MultiSphere's preprocessing in comparison to MultiSphere's K-best approach for a 6 user LDS-OFDM system employing 16 point codebooks. As we show in Fig. 20b the BER performance loss is negligible for an order of magnitude reduction in preprocessing complexity. Furthermore, the complexity of the method is adaptive to the SNR and channel.

VI. Conclusion

In this work we present g-MultiSphere; a massively parallel and near-optimal detection approach applicable to both well- and ill-determined non-orthogonal systems. The proposed approach can consistently provide substantial throughput gains in comparison to the existing detection approaches for MIMO, SEFDM and LDS-OFDM NOMA systems, with reduced complexity requirements. In particular, g-MultiSphere can efficiently support users more than twice the number of receiver antennas in a large multi-user MIMO environment while providing throughput gains of up to 60% in comparison to known approaches. By eliminating the need for sparse signal transmissions for NOMA schemes as specifically designed for MPA receiver processing, we show that g-MultiSphere can enable overloading factors beyond two with practical complexity and processing latency requirements. Due to this flexibility, g-MultiSphere can efficiently utilize the multiple access channel providing throughput gains while increasing the number of supported devices. In addition, g-MultiSphere can exploit the theoretical throughput gains of SEFDM systems.

Acknowledgment

The authors would like to thank the members of University of Surrey 5GIC (http://www.surrey.ac.uk/5GIC) for their support.

References

[1] C. Jayawardena and K. Nikitopoulos, "Massively parallel detection for non-orthogonal signal transmissions," in *Proc. IEEE Globecom Workshops*, Dec. 2018, pp. 1–6.

[2] R. Hoshyar, R. Razavi, and M. Al-Imari, "LDS-OFDM an efficient multiple access technique," in *Proc. IEEE VTC*, May 2010, pp. 1–5.

[3] K. Au *et al.*, "Uplink contention based SCMA for 5G radio access," in *Proc. IEEE GC Wkshps*, Dec. 2014, pp. 900–905.

[4] J. E. Mazo, "Faster-than-Nyquist signaling," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1451–1462, 1975.

[5] L. Landau, M. Dorpinghaus, and G. Fettweis, "Communications employing 1-bit quantization and oversampling at the receiver: Faster-than-nyquist signaling and sequence design," in *Proc. IEEE ICUWB*, Oct. 2015, pp. 1–5.

[6] F. Rusek and J. B. Anderson, "Constrained capacities for faster-than-Nyquist signaling," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 764–775, Feb. 2009.

[7] I. Kanaras, A. Chorti, M. R. D. Rodrigues, and I. Darwazeh, "Spectrally efficient FDM signals: Bandwidth gain at the expense of receiver complexity," in *Proc. IEEE ICC*, Jun. 2009, pp. 1–6.

[8] R. Razavi, M. Al-Imari, M. A. Imran, R. Hoshyar, and D. Chen, "On receiver design for uplink low density signature OFDM (LDS-OFDM)," *IEEE Trans. Commun.*, vol. 60, no. 11, pp. 3499–3508, Nov. 2012.

[9] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[10] C. Studer, A. Burg, and H. Bolcskei, "Soft-output sphere decoding: Algorithms and VLSI implementation," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 290–300, Feb. 2008.

[11] K. Nikitopoulos, J. Zhou, B. Congdon, and K. Jamieson, "Geosphere: Consistently turning MIMO capacity into throughput," in *Proc. ACM SIGCOMM*, Aug. 2014, pp. 631–642.

[12] L. G. Barbero and J. S. Thompson, "Extending a fixed-complexity sphere decoder to obtain likelihood information for turbo-MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 57, no. 5, pp. 2804–2814, Sep. 2008.

[13] Z. Guo and P. Nilsson, "Algorithm and implementation of the K-Best sphere decoding for MIMO detection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 491–503, Mar. 2006.

[14] S. Isam, I. Kanaras, and I. Darwazeh, "A truncated SVD approach for fixed complexity spectrally efficient FDM receivers," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2011, pp. 1584–1589.

[15] T. Xu and I. Darwazeh, "M-QAM signal detection for a non-orthogonal system using an improved fixed sphere decoder," in *Proc. CSNDSP*, Jul. 2014, pp. 623–627.

[16] T. Xu and I. Darwazeh, "A soft detector for spectrally efficient systems with non-orthogonal overlapped sub-carriers," *IEEE Commun. Lett.*, vol. 18, no. 10, pp. 1847–1850, Oct. 2014.

[17] H. Huawei, *Candidate Schemes for Superposition Transmission*, document R1-152493, 3GPP, May 2015.

[18] P. Popovski *et al.*, "Requirement analysis and design approaches for 5G air interface," METIS Project, EU FP7, Brussels, Belgium, Tech. Rep. ICT-317669-METIS/D2.1, Aug. 2013.

[19] R. Courtland, "Transistors could stop shrinking in 2021," *IEEE Spectr.*, vol. 53, no. 9, pp. 9–11, Sep. 2016.

[20] K. Nikitopoulos, G. Georgis, C. Jayawardena, D. Chatzipanagiotis, and R. Tafazolli, "Massively parallel tree search for high-dimensional sphere decoders," *IEEE Trans. Parallel Distrib. Syst*, vol. 30, no. 10, pp. 2309–2325, Oct. 2019.

[21] C. Husmann, G. Georgis, K. Nikitopoulos, and K. Jamieson, "FlexCore: Massively parallel and flexible processing for large MIMO access points," in *Proc. USENIX NSDI*, 2017, pp. 197–211.

[22] A. Chorti, I. Kanaras, M. R. D. Rodrigues, and I. Darwazeh, "Joint channel equalization and detection of spectrally efficient FDM signals," in *Proc. IEEE PIMRC*, Sep. 2010, pp. 177–182.

[23] J. Bao, Z. Ma, Z. Ding, G. K. Karagiannidis, and Z. Zhu, "On the design of multiuser codebooks for uplink SCMA systems," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 1920–1923, Oct. 2016.

[24] P. C. Hansen, "Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank," *SIAM J. Sci. Statist. Comput.*, vol. 11, no. 3, pp. 503–518, 1990.

[25] T. Cui and C. Tellambura, "An efficient generalized sphere decoder for rank-deficient MIMO systems," in *Proc. IEEE 60th Veh. Technol. Conf., VTC2004-Fall*, vol. 5, Sep. 2004, pp. 3689–3693.

[26] D. Wubben, R. Bohnke, V. Kuhn, and K.-D. Kammeyer, "MMSE extension of V-BLAST based on sorted QR decomposition," in *Proc. IEEE Veh. Technol. Conf. (VTC-Fall)*, vol. 1. Oct. 2003, pp. 508–512.

[27] A. Burg, M. Borgmann, M. Wenk, M. Zellweger, W. Fichtner, and H. Bolcskei, "VLSI implementation of MIMO detection using the sphere decoding algorithm," *IEEE J. Solid-State Circuits*, vol. 40, no. 7, pp. 1566–1577, Jul. 2005.

[28] K. Nikitopoulos, D. Chatzipanagiotis, C. Jayawardena, and R. Tafazolli, "MultiSphere: Massively parallel tree search for large sphere decoders," in *Proc. IEEE GLOBECOM*, Dec. 2016, pp. 1–6.

[29] C.-H. Liao *et al.*, "Combining orthogonalized partial metrics: Efficient enumeration for soft-input sphere decoder," in *Proc. IEEE PIMRC*, Sep. 2009, pp. 1287–1291.

[30] K. Nikitopoulos, D. Zhang, I.-W. Lai, and G. Ascheid, "Complexity-efficient enumeration techniques for soft-input, soft-output sphere decoding," *IEEE Commun. Lett.*, vol. 14, no. 4, pp. 312–314, Apr. 2010.

[31] K. Nikitopoulos and G. Ascheid, "Approximate MIMO iterative processing with adjustable complexity requirements," *IEEE Trans. Veh. Technol.*, vol. 61, no. 2, pp. 639–650, Feb. 2012.

[32] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 2002.

[33] J. Jalden, L. G. Barbero, B. Ottersten, and J. S. Thompson, "Full diversity detection in MIMO systems with a fixed-complexity sphere decoder," in *Proc. IEEE ICASSP*, Apr. 2007, pp. III-49–III-52.

Chathura Jayawardena received the B.Eng. degree in electronic engineering and the M.Sc. degree in mobile communications from the University of Surrey, Guildford, U.K., in 2014 and 2015, respectively, where he is currently pursuing the Ph.D. degree in electronic engineering with the Institute for Communication Systems. His research interests include signal processing for communications, with an emphasis on detection methods for non-orthogonal transmission schemes.

Konstantinos Nikitopoulos (M'07) is currently an Associate Professor with the Electrical and Electronic Engineering Department, Institute for Communication Systems, University of Surrey, Guildford, U.K., and a member of the 5G Innovation Centre. He is the Director of the Wireless Systems Lab, Institute for Communication Systems, and leads the Proof-of-Concept and mmWave Solutions Work Area in 5GIC. He has held research positions with RWTH Aachen University, the University of California at Irvine, Irvine, and the University College London. He was also a consultant for the Hellenic General Secretariat for Research and Technology, where he also served as a National Delegate of Greece to the Joint Board on Communication Satellite Programs of European Space Agency. He was a recipient of the prestigious First Grant of the U.K.'s Engineering and Physical Sciences Research Council.