Mansour *EURASIP Journal on Wireless Communications and Networking* (2015) 2015:95
DOI 10.1186/s13638-015-0274-9

EURASIP Journal on
Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

A low-complexity MIMO subspace detection algorithm

Mohammad M Mansour

Abstract

A low-complexity multiple-input multiple-output (MIMO) subspace detection algorithm is proposed. It is based on decomposing a MIMO channel into multiple subsets of decoupled streams that can be detected separately. The new scheme employs triangular decomposition followed by elementary matrix operations to transform the channel into a generalized elementary matrix whose structure matches the subsets of streams to be detected. The proposed approach avoids matrix inversion and allows subsets to overlap, thus achieving better diversity gain. An optimized detector architecture based on a 2-by-2 ML detector core is also presented. Simulations demonstrate that the proposed algorithm performs to within a few tenths of a dB from the optimum detection algorithm.

Keywords: Maximum likelihood; MIMO detection; LLR; Subspace detection

1 Introduction

With the advent of smart mobile devices, the demand for wireless access to broadband networks has been rapidly increasing in the past decade. Service providers are constantly faced with a major challenge of meeting contradictory requirements for higher data rates, improved quality of service (QoS), and better network capacity, while maintaining the transmit power and bandwidth budgets. To achieve these targets, novel enabling technologies need to be considered.

Exploiting the spatial dimension by using multiple-input multiple-output (MIMO) antenna systems is one of the key-enabling technologies for achieving high spectral efficiency in modern wireless communications standards. MIMO technology improves both the spectral efficiency and the QoS of wireless communication systems. However, detection of spatially multiplexed MIMO streams plays a key role in receiver design both in terms of performance and complexity [1] and has remained an active area of research. Schemes for MIMO detection are either based on joint-stream detection to separate the streams, such as maximum likelihood (ML) detection [2-4], or subset-stream detection such as those employed in successive interference cancellation (SIC), interference coordination, transmit beamforming, network MIMO, and coordinated multi-point transmission and reception [5-9].

A plethora of MIMO detectors have appeared in the literature on this subject, offering various performance-complexity tradeoffs. Suboptimal zero-forcing (ZF) and minimum mean-squared error (MMSE) detectors [10], as well as the nonlinear parallel and successive interference cancellation schemes [11-14], require relatively low complexity but sacrifice performance. On the other hand, tree-search or list-based detectors require substantially higher complexity but can offer (near-)ML performance such as the well-known sphere decoding algorithm [2-4,15-19]. Other tree-search schemes, such as the K-Best algorithm [20-26], address the non-deterministic throughput aspects of sphere decoders. Practical implementation aspects have been investigated in [18,23,25-39].

Subspace detection based on channel decomposition offers a good compromise between performance and complexity. In [6,7], a method was presented in which the effective MIMO channel matrix **H** is uniformly decomposed into identical parallel subchannels using geometric mean decomposition (GMD). In [8], a related scheme was presented in which **H** is block-wise diagonalized to narrow down the number of jointly detected streams to two, and then GMD is applied to balance capacity within each pair of subchannels. The scheme was generalized in [9] to allow joint detection of several overlapping subchannels using QR decomposition (QRD). By allowing subspaces

Correspondence: mmansour@ieee.org
American University of Beirut, Bliss Street, 11-0236 Beirut, Lebanon

to overlap, additional diversity can be gathered by putting a low reliable data stream into several detection sets. Other subspace methods employ projection operators and lists to generate candidates for interference cancellation in equalization schemes (e.g., [40-44]).

The LORD algorithm proposed in [45,46] can be viewed as a special class of subspace MIMO detectors. It achieves ML performance (in the Max-log-MAP [47] sense) on two transmit antennas, but its performance degrades when the number of antennas increases. In [48], the LORD algorithm was generalized to four transmit antennas by using matrix inversion to decompose $\mathbf{H}$ into single streams.

Contributions: In this paper, we propose an efficient near-ML soft-output MIMO detection algorithm that jointly detects subsets of decoupled streams by transforming $\mathbf{H}$ into a generalized elementary matrix. A matrix transformation is analytically derived to induce the desired structure on $\mathbf{H}$, while avoiding computationally complex operations such as pseudo-inversion. This is achieved by applying QL decomposition followed by elementary matrix operations on $\mathbf{H}$ in a manner analogous to the modified QL decomposition algorithm based on the Gram-Schmidt orthogonalization procedure, thus avoiding the need for expensive matrix inversion operations. Various decomposition structures are investigated, including the option of allowing the decomposition sets to overlap resulting in additional diversity gain. Furthermore, we show that a parallel MIMO detector can be constructed from 2 × 2 component detector cores that decouple the streams in the subsets in parallel. Finally, we show that for two streams, the proposed algorithm is optimal and reduces to the LORD algorithm [45,46]. When the subsets include single streams only, the algorithm reduces to that of [48]. The advantages of the proposed algorithm in attaining near-ML performance outperforming that of the sphere decoder [30] and the near-ML algorithm of [48] are demonstrated through computer simulations of the bit error-rate of a MIMO system.

The rest of the paper is organized as follows. Section 2 introduces the system model, and Section 3 reviews ML detection for two streams. Sections 4 and 5 present the proposed matrix decomposition scheme and a simplified construction using QLD. The detection algorithm is detailed in Section 6. Section 7 presents simulation results, while Section 8 ends the paper with concluding remarks.

2 System model

Consider a MIMO system with N transmit (Tx) antennas and $M \geq N$ receive (Rx) antennas. Assuming perfect channel knowledge at the receiver, the equivalent complex baseband input-output system relation can be modeled as $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$, where $\mathbf{y} \in \mathcal{C}^{M\times 1}$ is the received complex signal vector, $\mathbf{H} \in \mathcal{C}^{M\times N}$ is the complex channel matrix, and $\mathbf{x} = [x_1\ x_2 \cdots x_N]^T \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$ is the $N\times 1$ transmitted complex symbol vector. Each symbol x_n belongs to a complex constellation $\mathcal{X}_n$ of size $Q_n = 2^{q_n}$ and normalized so that $\mathbb{E}\left[x_n^* x_n\right] = 1$, and is formed from a set of q_n coded bit-interleaved sequence $\mathbf{b}_n = \left(b_{n,1}, b_{n,2}, \cdots, b_{n,q_n}\right)$ over the binary field $\mathcal{F}_2$. The effect of thermal noise is modeled as a zero-mean complex Gaussian circularly symmetric random vector $\mathbf{n} \in \mathcal{C}^{M\times 1}$ with covariance matrix $\mathbb{E}[\mathbf{n}\mathbf{n}^*] = \sigma_{\mathbf{n}}^2 \mathbf{I}_M$. The signal-to-noise ratio (SNR) is defined as $\mathrm{SNR} = N/\sigma_{\mathbf{n}}^2$. $\mathbb{E}[\cdot]$ stands for the expected value, $(\cdot)^T$ and $(\cdot)^*$ for the transpose and conjugate transpose, and $\mathbf{I}_M$ for the $M \times M$ identity matrix.

Assuming equiprobable symbols, ML MIMO detection algorithms achieve optimum performance by finding the symbol vector $\mathbf{x} \in \mathcal{X}$ that is closest to the received vector $\mathbf{y}$ under the Euclidean distance metric

$$d\left(\mathbf{x}\right) \triangleq \left\|\mathbf{y} - \mathbf{H}\mathbf{x}\right\|^2 = \left\|\tilde{\mathbf{y}} - \mathbf{L}\mathbf{x}\right\|^2, \tag{1}$$

where $\mathbf{H} = \mathbf{Q}\mathbf{L}$ is the QL decomposition (QLD) [49] of $\mathbf{H}$ into a unitary matrix $\mathbf{Q} \in \mathcal{C}^{M\times N}$ and a lower triangular matrix (LTM) $\mathbf{L} \in \mathcal{C}^{N\times N}$ with positive real diagonal elements, and $\tilde{\mathbf{y}} = \mathbf{Q}^*\mathbf{y} = \mathbf{L}\mathbf{x} + \mathbf{Q}^*\mathbf{n} \in \mathcal{C}^{N\times 1}$ denotes the transformed received signal vector. Since $\mathbf{Q}$ is unitary, it preserves Euclidean norm as well as noise statistics. Hence, for equiprobable symbols, a 'hard-decision' ML MIMO detector finds $\mathbf{x}^{\mathrm{ML}} \in \mathcal{X}$ such that $\mathbf{H}\mathbf{x}^{\mathrm{ML}}$ is closest to $\mathbf{y}$ in $\mathcal{C}^{M\times 1}$ (or equivalently, $\mathbf{L}\mathbf{x}^{\mathrm{ML}}$ is closest to $\tilde{\mathbf{y}}$ in $\mathcal{C}^{N\times 1}$). This is essentially an integer least-squares problem [4] of the form

$$d^{\mathrm{ML}} = \min_{\mathbf{x}\in\mathcal{X}} \left\|\tilde{\mathbf{y}} - \mathbf{L}\mathbf{x}\right\|^2 \tag{2}$$

$$\mathbf{x}^{\mathrm{ML}} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\|\tilde{\mathbf{y}} - \mathbf{L}\mathbf{x}\right\|^2. \tag{3}$$

In MIMO systems employing soft-input channel decoders however, ML MIMO detectors generate soft-outputs (SO) in the form of log-likelihood ratios (LLRs) by searching for other 'closest' symbol vectors to $\tilde{\mathbf{y}}$. The (unscaled) LLR value associated with bit $b_{n,k}$ is given by

$$\Lambda_{n,k}^{\mathrm{ML}} = \min_{\mathbf{x}\in\mathcal{X}_{n,k}^{(0)}} d(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}_{n,k}^{(1)}} d(\mathbf{x}),\ n = 1, \cdots, N;\ k = 1, \cdots, q_n, \tag{4}$$

where $\mathcal{X}_{n,k}^{(0)} = \left\{\mathbf{x} \in \mathcal{X} : b_{n,k} = 0\right\}$ and $\mathcal{X}_{n,k}^{(1)} = \{\mathbf{x} \in \mathcal{X} : b_{n,k} = 1\}$ are the subsets of symbol vectors in $\mathcal{X}$ that have their corresponding kth bit in the nth symbol 0 and 1, respectively.

3 Optimum 2 × 2 soft-output MIMO detection

In general, finding the ML solution requires computing $\prod_{n=1}^{N} Q_n$ distance metrics. When $N = 2$, a simplification [45] can be applied to reduce the number of computations from $Q_1 \cdot Q_2$ to $Q_1 + Q_2$ by triangularizing the channel matrix as $\mathbf{H} = \mathbf{Q}_1\mathbf{L}_1$, with $\mathbf{Q}_1$ being unitary and $\mathbf{L}_1$ being a LTM, leading to:

$$\mathbf{y}-\mathbf{H}\mathbf{x}\longrightarrow\begin{bmatrix}\tilde{y}_1\\ \tilde{y}_2\end{bmatrix}-\begin{bmatrix}a_1 & 0\\ c_1 & b_1\end{bmatrix}\cdot\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\tilde{\mathbf{y}}_1-\mathbf{L}_1\mathbf{x},\quad(5)$$

where $\tilde{\mathbf{y}}_1=\mathbf{Q}_1^*\mathbf{y}$, with $a_1,b_1\in\mathcal{R}^+$ and $c_1\in\mathcal{C}$. Then

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}}\left\|\tilde{\mathbf{y}}_1-\mathbf{L}_1\mathbf{x}\right\|^2&=\min_{\substack{x_1\in\mathcal{X}_1\\ x_2\in\mathcal{X}_2}}\left(\left|\tilde{y}_1-a_1x_1\right|^2+\left|\tilde{y}_2-c_1x_1-b_1x_2\right|^2\right)\\&=\min_{x_1\in\mathcal{X}_1}\left(\left|\tilde{y}_1-a_1x_1\right|^2+\min_{x_2\in\mathcal{X}_2}\left|\tilde{y}_2-c_1x_1-b_1x_2\right|^2\right)\\&=\min_{x_1\in\mathcal{X}_1}\left(\left|\tilde{y}_1-a_1x_1\right|^2+\left|\tilde{y}_2-c_1x_1-b_1\hat{x}_2\right|^2\right)\\&\triangleq\min_{x_1\in\mathcal{X}_1}d_1(x_1)\qquad(6)\end{aligned}$$

where $\hat{x}_2$ is obtained by slicing $(\tilde{y}_2-c_1x_1)/b_1\in\mathcal{C}$ to the nearest constellation point in $\mathcal{X}_2$ using the operator $\lfloor\alpha\rceil_{\mathcal{X}_n}\triangleq\arg\min_{x\in\mathcal{X}_n}|\alpha-x|$:

$$\hat{x}_2=\left\lfloor(\tilde{y}_2-c_1x_1)/b_1\right\rceil_{\mathcal{X}_2}\in\mathcal{X}_2.\qquad(7)$$

Hence (6) requires only $|\mathcal{X}_1|=Q_1$ distance computations. The LLRs of the bits in symbol x_1 can simply be obtained as

$$\Lambda^{\mathrm{ML}}_{1,k}=\min_{x_1\in\mathcal{X}^{(0)}_{1,k}}d_1(x_1)-\min_{x_1\in\mathcal{X}^{(1)}_{1,k}}d_1(x_1),\quad k=1,\cdots,q_1.\quad(8)$$

To obtain the LLRs of the bits in x_2 however, we triangularize $\mathbf{H}$ as $\mathbf{Q}_2\mathbf{L}_2$ so that a zero appears in the upper left corner:

$$\mathbf{y}-\mathbf{H}\mathbf{x}\longrightarrow\begin{bmatrix}\bar{y}_1\\ \bar{y}_2\end{bmatrix}-\begin{bmatrix}0 & a_2\\ b_2 & c_2\end{bmatrix}\cdot\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\bar{\mathbf{y}}_2-\mathbf{L}_2\mathbf{x},\quad(9)$$

where now $\bar{\mathbf{y}}_2=\mathbf{Q}_2^*\mathbf{y}$, and $a_2,b_2\in\mathcal{R}^+$ and $c_2\in\mathcal{C}$. Then

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}}\left\|\bar{\mathbf{y}}_2-\mathbf{L}_2\mathbf{x}\right\|^2&=\min_{\substack{x_1\in\mathcal{X}_1\\ x_2\in\mathcal{X}_2}}\left(\left|\bar{y}_1-a_2x_2\right|^2+\left|\bar{y}_2-c_2x_2-b_2x_1\right|^2\right)\\&=\min_{x_2\in\mathcal{X}_2}\left(\left|\bar{y}_1-a_2x_2\right|^2+\left|\bar{y}_2-c_2x_2-b_2\hat{x}_1\right|^2\right)\\&\triangleq\min_{x_2\in\mathcal{X}_2}d_2(x_2)\qquad(10)\end{aligned}$$

where $\hat{x}_1=\left\lfloor(\bar{y}_2-c_2x_2)/b_2\right\rceil_{\mathcal{X}_1}$. The LLRs of the bits in x_2 are given by

$$\Lambda^{\mathrm{ML}}_{2,k}=\min_{x_2\in\mathcal{X}^{(0)}_{2,k}}d_2(x_2)-\min_{x_2\in\mathcal{X}^{(1)}_{2,k}}d_2(x_2),\quad k=1,\cdots,q_2.\quad(11)$$

Since $\mathbf{Q}_1,\mathbf{Q}_2$ are unitary, the ML solutions in (6), (10) are identical. To find the hard-decision ML solution, only one-sided QLD is needed on either layer 1 or 2. A list of Q_n distances $\{d_n(x),\ \forall x\in\mathcal{X}_n\}$ is generated by enumerating all symbols $x\in\mathcal{X}_n$, $n=1$ or $n=2$, and the minimum is selected. However, to generate soft LLRs, two-sided decompositions are needed, and *two* lists of distances for $n=1$ *and* $n=2$ must be computed to select the appropriate minima in (8) and (11).

4 Extensions to higher-order layers

The previous optimization cannot be extended in a straightforward manner to $N=3$ or more layers because the structure of the lower triangular matrix $\mathbf{L}$ includes off-diagonal terms that prevent searching for the ML solution by enumerating symbols on one layer and finding the minima through slicing individually on all other layers in parallel. More specifically, in Figure 1a, the presence of the de-marked entries in the LTM implies that determining the ML solution requires enumerating symbols on $N-1$ layers and slicing only on the last layer, as is typically done in tree-based detectors (e.g., [30]), and hence still requiring $O\left(\prod_n Q_n\right)$ complexity rather than $O\left(\sum_n Q_n\right)$.

One desirable structure of $\mathbf{H}$ for a four-layer MIMO system would be as shown in Figure 1b, in which the red-marked entries are zeroed-out. Here, by enumerating symbols on layer 1, the minimum distances and associated symbols on layers 2 to 4 can be searched for in parallel through slicing only on the corresponding layers, similar to the two-layer system. This suffices to compute the LLRs associated with the bits on layer-1 symbol. A similar process is repeated by decomposing $\mathbf{H}$ according to the structures shown in Figure 1c,d,e [48] to compute the LLRs for bits associated with layers 2 to 4.

Other 'punctured' structures are also possible for a 4×4 system as illustrated in Figure 2. They differ in 1) the number of layers over which symbols are enumerated (*enumeration set*), 2) the submatrix structure used to propagate these enumerated symbols and cancel their interference effect from the remaining layers (*interference cancellation set*), and 3) the number of layers in which the minimum distance and associated symbol can be obtained by slicing after interference cancellation (*slicer set*). Let E denote the size of the enumeration set, S the size of the slicer set, and

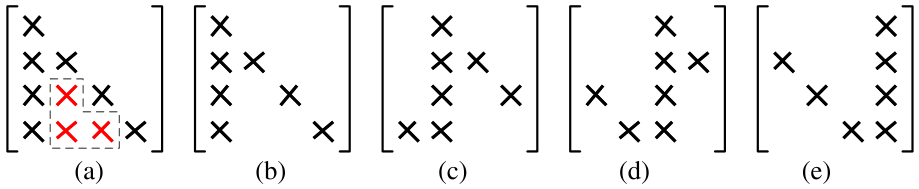

Figure 1 4 × 4 channel matrix structures: (a) full; and (b-e) punctured structures for every layer.

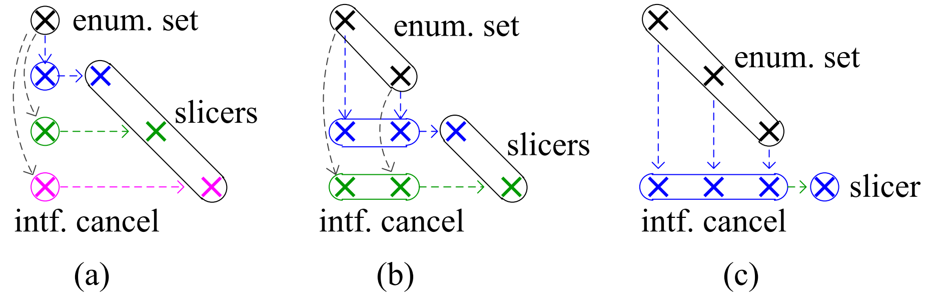

Figure 2 $(1, 3 \times 1, 3)$ **(a),** $(2, 2 \times 2, 2)$ **(b), and** $(3, 1 \times 3, 1)$ **(c) punctured structures.**

$S \times E$ the size of the interference cancellation set. We refer to this structure using the triplet $(E, S \times E, S)$. For example, in Figure 2a, we enumerate over $E = 1$ layer only, cancel interference from this layer to the three other layers using a 3×1 structure, and slice over $S = 3$ layers. In the structure in Figure 2b, we enumerate over $E = 2$ layers, cancel interference using a 2×2 structure, and slice over $S = 2$ layers.

LLR values are generated for bits in symbols included in the enumeration set only. Complementary structures that enumerate symbols on other layers are required to generate their respective LLRs. For example, the $(1, 3 \times 1, 3)$ structure requires three other similar structures to generate LLRs for layers 2 to 4 (see Figure 1c,d,e). The $(2, 2 \times 2, 2)$ structure of Figure 2b on the other hand requires only one identical structure to generate LLRs for both layers 3 and 4, while Figure 2c requires one non-identical $(1, 3 \times 1, 3)$ structure to handle layer 4.

4.1 Preliminaries

Let $\mathbf{H} = [\mathbf{h}_1\ \mathbf{h}_2\ \cdots\ \mathbf{h}_N]$, and assume $\mathbf{H}$ has full column rank. For simplicity, we assume $N = M$ in the remainder of this work. We seek a matrix $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_N] \in \mathcal{C}^{N \times N}$ that transforms $\mathbf{H}$ into a *punctured* LTM $\mathbf{L} =[l_{ij}] \in \mathcal{C}^{N \times N}$ with $l_{ii} \in \mathcal{R}^+$, such that:

$$\mathbf{W}^*\mathbf{H} = \mathbf{L}. \tag{12}$$

The aim is to induce a specific pattern of zeros below the main diagonal of $\mathbf{L}$ by appropriately choosing the columns of $\mathbf{W}$. Setting $\mathbf{W} = (\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*$ to be the left Moore-Penrose pseudo-inverse of $\mathbf{H}$ would obviously zero-out all entries in $\mathbf{L}$ below the main diagonal, resulting in $\mathbf{L} = \mathbf{I}_N$. On the other hand, choosing $\mathbf{W}$ to be an orthonormal basis of the column space of $\mathbf{H}$, would transform $\mathbf{H}$ into a regular (i.e., unpunctured) LTM, with $\mathbf{W}$ being unitary.

In general, if $\mathbf{L}$ is punctured, then $\mathbf{W}$ is non-unitary. However, we impose the condition on the column vectors of $\mathbf{W}$ to have unit length, i.e., $\mathbf{w}_n^*\mathbf{w}_n = 1$ for $n = 1, \cdots, N$, or

$$\operatorname{diag}\left(\mathbf{W}^*\mathbf{W}\right) =[1\ 1\ \cdots\ 1]^T_{1 \times N}. \tag{13}$$

This guarantees that, when $\mathbf{y}$ is left-multiplied by $\mathbf{W}^*$, the transformed noise vector $\mathbf{W}^*\mathbf{n}$ has an unaltered covariance matrix $\mathbb{E}[\mathbf{W}^*\mathbf{nn}^*\mathbf{W}] = \sigma_{\mathbf{n}}^2\mathbf{I}_N$. Note also that non-zero off-diagonal entries of the Gram matrix $\mathbf{W}^*\mathbf{W}$ correspond to pairs of column vectors in $\mathbf{W}$ that are non-orthogonal.

4.2 Proposed WL decomposition scheme

Let $\mathbf{P} = \mathbf{H}(\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*$ be the orthogonal projection onto the column space of $\mathbf{H}$, and $\mathbf{P}^\perp = \mathbf{I} - \mathbf{H}(\mathbf{H}^*\mathbf{H})^{-1}\mathbf{H}^*$ be the orthogonal projection onto the left nullspace of $\mathbf{H}$. Let $\mathbf{H}_\mathcal{I}$ be the submatrix formed by the columns of $\mathbf{H}$ whose index n belongs to set $\mathcal{I}$. For example, if $\mathbf{H} = [\mathbf{h}_1\ \mathbf{h}_2\ \mathbf{h}_3\ \mathbf{h}_4]$ and $\mathcal{I} = \{1, 2\}$, then $\mathbf{H}_\mathcal{I} =[\mathbf{h}_1\ \mathbf{h}_2]$.

Let $\mathcal{I}_n$ be column index set of the entries in the nth row of $\mathbf{H}$ to be zeroed out. Define the nth column vector $\widetilde{\mathbf{w}}_n = \mathbf{P}^\perp_{\mathcal{I}_n}\mathbf{h}_n$, where

$$\mathbf{P}^\perp_{\mathcal{I}_n} = \mathbf{I}_N - \mathbf{H}_{\mathcal{I}_n}\left(\mathbf{H}^*_{\mathcal{I}_n}\mathbf{H}_{\mathcal{I}_n}\right)^{-1}\mathbf{H}^*_{\mathcal{I}_n} \tag{14}$$

and $\mathbf{H}_{\mathcal{I}_n} = \{\mathbf{h}_m \mid m \in \mathcal{I}_n\}$. To satisfy (13), first note that

$$\widetilde{\mathbf{w}}_n^*\widetilde{\mathbf{w}}_n = \mathbf{h}_n^*\left(\mathbf{P}^\perp_{\mathcal{I}_n}\right)^*\mathbf{P}^\perp_{\mathcal{I}_n}\mathbf{h}_n = \mathbf{h}_n^*\mathbf{P}^\perp_{\mathcal{I}_n}\mathbf{h}_n = \widetilde{\mathbf{w}}_n^*\mathbf{h}_n,$$

where the second equality follows since $\mathbf{P}^\perp_{\mathcal{I}_n}$ is a projection matrix, and hence $\left(\mathbf{P}^\perp_{\mathcal{I}_n}\right)^* = \mathbf{P}^\perp_{\mathcal{I}_n}$ and $\mathbf{P}^\perp_{\mathcal{I}_n}\mathbf{P}^\perp_{\mathcal{I}_n} = \mathbf{P}^\perp_{\mathcal{I}_n}$. Hence the normalized vector $\mathbf{w}_n$, defined as

$$\mathbf{w}_n = \frac{\widetilde{\mathbf{w}}_n}{\|\widetilde{\mathbf{w}}_n\|},$$

with $\|\widetilde{\mathbf{w}}_n\| = \sqrt{\mathbf{h}_n^*\mathbf{P}^\perp_{\mathcal{I}_n}\mathbf{h}_n}$, has unit length.

We have the following main result:

Theorem 1 (WL decomposition). *Let $\mathcal{I}_n, n = 1, \cdots, N$, be the column index sets where puncturing is desired in each row n of* $\mathbf{H}$. *Form the submatrices $\mathbf{H}_{\mathcal{I}_n}$ and their projection matrices $\mathbf{P}^\perp_{\mathcal{I}_n}$ as noted above. Let* $\mathbf{D} =[d_n] \in \mathcal{R}^+$ *be a diagonal matrix whose entries are given by $d_n = 1\Big/\sqrt{\mathbf{h}_n^*\mathbf{P}^\perp_{\mathcal{I}_n}\mathbf{h}_n}$, $n = 1, \cdots, N$. Then the matrix*

$$\mathbf{W}^* = \mathbf{D}\begin{bmatrix} \mathbf{h}_1^*\mathbf{P}_{\mathcal{I}_1}^{\perp} \\ \mathbf{h}_2^*\mathbf{P}_{\mathcal{I}_2}^{\perp} \\ \vdots \\ \mathbf{h}_N^*\mathbf{P}_{\mathcal{I}_N}^{\perp} \end{bmatrix}, \tag{15}$$

when right multiplied by $\mathbf{H}$*, zeros out the entries in the rows of* $\mathbf{H}$ *at column positions given in* $\mathcal{I}_n$*, for all n, and satisfies condition (13).*

Proof. Consider the product of $\mathbf{w}_n^*$ with $\mathbf{H}_{\mathcal{I}_n}$:

$$\begin{aligned} \mathbf{w}_n^*\mathbf{H}_{\mathcal{I}_n} &= d_n\mathbf{h}_n^*\left(\mathbf{I}_N - \mathbf{H}_{\mathcal{I}_n}(\mathbf{H}_{\mathcal{I}_n}^*\mathbf{H}_{\mathcal{I}_n})^{-1}\mathbf{H}_{\mathcal{I}_n}^*\right)\mathbf{H}_{\mathcal{I}_n} \\ &= d_n\mathbf{h}_n^*\left(\mathbf{H}_{\mathcal{I}_n} - \mathbf{H}_{\mathcal{I}_n}\right) = \mathbf{0}_{1\times|\mathbf{H}_{\mathcal{I}_n}|} \end{aligned}$$

Also, $\mathbf{w}_n^*\mathbf{h}_m \neq 0$ for all $m \notin \mathcal{I}_n$, $1 \leq m \leq N$. Hence $\mathbf{w}_n$ nulls the nth row of $\mathbf{H}$ only at the column indices given in $\mathcal{I}_n$. □

Example 1. *For a* 4×4 *MIMO system, choosing the puncturing sets as* $\mathcal{I}_1 = \{2,3,4\}, \mathcal{I}_2 = \{3,4\}, \mathcal{I}_3 = \{2,4\}, \mathcal{I}_4 = \{2,3\}$*, transforms* $\mathbf{H}$ *into a LTM* $\mathbf{L}$ *with entries* $l_{32} = l_{42} = l_{43} = 0$ *as follows (Figure 2a):*

$$\mathbf{W}^*\mathbf{H} = \mathbf{D}\begin{bmatrix} \mathbf{h}_1^*\mathbf{P}_{2,3,4}^{\perp} \\ \mathbf{h}_2^*\mathbf{P}_{3,4}^{\perp} \\ \mathbf{h}_3^*\mathbf{P}_{2,4}^{\perp} \\ \mathbf{h}_4^*\mathbf{P}_{2,3}^{\perp} \end{bmatrix}\mathbf{H} = \begin{bmatrix} \times & & & \\ \times & \times & & \\ \times & & \times & \\ \times & & & \times \end{bmatrix} = \mathbf{L}$$

Choosing $\mathcal{I}_1 = \{2,3,4\}$, $\mathcal{I}_2 = \{1,3,4\}$, $\mathcal{I}_3 = \{4\}$, $\mathcal{I}_4 = \{3\}$, *results in the form of Figure 2b, while the choice* $\mathcal{I}_1 = \{2,3,4\}$, $\mathcal{I}_2 = \{1,3,4\}$, $\mathcal{I}_3 = \{1,2,4\}$, $\mathcal{I}_4 = \{\varnothing\}$ *results in the form of Figure 2c.* □

The $\mathbf{W}, \mathbf{L}$ pair in (12) that punctures $\mathbf{H}$ into a given lower triangular structure is not unique since $\mathbf{W}$ is non-unitary. But any other pair that generates the same structure is related to $\mathbf{W}, \mathbf{L}$ by a matrix transformation as the next lemma shows.

Lemma 1 (Similar WL decompositions)**.** *Let* $\mathbf{W}_1, \mathbf{W}_2$ *be two distinct matrices that transform* $\mathbf{H}$ *into two distinct but identically punctured LTMs* $\mathbf{L}_1$ *and* $\mathbf{L}_2$*. Then the orthonormal bases of the column space of* $\mathbf{W}_1$ *and* $\mathbf{W}_2$ *are both identical to that of* $\mathbf{H}$*, and*

$$\mathbf{L}_1 = \mathbf{R}_1^*\left(\mathbf{R}_2^*\right)^{-1}\mathbf{L}_2, \quad \mathbf{W}_1 = \mathbf{R}_1\mathbf{R}_2^{-1}\mathbf{W}_2 \tag{16}$$

Proof. We have $\mathbf{W}_1^*\mathbf{H} = \mathbf{L}_1$ and $\mathbf{W}_2^*\mathbf{H} = \mathbf{L}_2$. Let $\mathbf{W}_1 = \mathbf{Q}_1\mathbf{R}_1$ and $\mathbf{W}_2 = \mathbf{Q}_2\mathbf{R}_2$ denote the QR decompositions of $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively. Then $\mathbf{H} = \mathbf{Q}_1\left(\mathbf{R}_1^*\right)^{-1}\mathbf{L}_1 = \mathbf{Q}_2\left(\mathbf{R}_2^*\right)^{-1}\mathbf{L}_2$ where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are unitary, with both $\left(\mathbf{R}_1^*\right)^{-1}\mathbf{L}_1$ and $\left(\mathbf{R}_2^*\right)^{-1}\mathbf{L}_2$ being lower triangular matrices. But $\mathbf{H}$ admits a unique QLD in the form $\mathbf{H} = \mathbf{QL}$, hence $\mathbf{Q}_1 = \mathbf{Q}_2 = \mathbf{Q}$ and $\left(\mathbf{R}_1^*\right)^{-1}\mathbf{L}_1 = \left(\mathbf{R}_2^*\right)^{-1}\mathbf{L}_2$. □

5 Reduced-complexity WL decomposition using QLD

A brute force approach for computing $\mathbf{W}$ involves extensive matrix inversion, which is computationally expensive and prone to numerical error when executed on DSP or dedicated hardware with finite precision. We develop next an efficient scheme to determine $\mathbf{W}$ using the modified Gram-Schmidt orthogonalization procedure [49-51] followed by elementary matrix operations. From Lemma 1, any other $\mathbf{W}$ that produces an identical structure is related by (16).

Assume that $\mathbf{H}$ is decomposed into a unitary matrix $\mathbf{Q}_1 = [\mathbf{q}_1\ \mathbf{q}_2\ \cdots\ \mathbf{q}_N]$ and a lower triangular matrix $\mathbf{L}_1 = [\,l_{ij}]_{N\times N}$ with real and positive diagonal elements. We then have $\mathbf{Q}_1^*\mathbf{H} = \mathbf{L}_1$. Obviously, $\mathbf{q}_1^*\mathbf{q}_1 = 1$ and $\mathbf{q}_1^*\mathbf{h}_m = 0$ for all $m = 2, \cdots, N$. Hence, $\mathbf{w}_1 = \mathbf{q}_1$. Now consider row $1 < n \leq N$ of $\mathbf{L}_1$, and assume the mth entry l_{nm}, $m < n$, is to be nulled. We have $\mathbf{q}_n^*\mathbf{h}_m = l_{nm} \in \mathcal{C}$ and $\mathbf{q}_m^*\mathbf{h}_m = l_{mm} \in \mathcal{R}^+$, from which it follows that $\left(\mathbf{q}_n^* - \mathbf{q}_m^*\frac{l_{nm}}{l_{mm}}\right)\mathbf{h}_m = 0$. Therefore, the matrix operations

$$\mathbf{q}_n = \mathbf{q}_n - \mathbf{q}_m l_{nm}^*/l_{mm} \tag{17}$$

$$l_{nj} = l_{nj} - l_{mj}l_{nm}/l_{mm}, \text{ for } j = 1, \cdots, m, \tag{18}$$

puncture the required entry and update the columns of $\mathbf{Q}_1$ accordingly. These operations are repeated for all other column entries $m < n$ to be punctured in row n. Finally, $\mathbf{q}_n$ is normalized to have unit length, and the non-zero entries in row n of $\mathbf{L}_1$ are updated:

$$l_{nj} = l_{nj}/\|\mathbf{q}_n\|\ , \text{ for } j = 1, \cdots, n \tag{19}$$

$$\mathbf{q}_n = \mathbf{q}_n/\|\mathbf{q}_n\| \tag{20}$$

The operations in (17)-(18) followed by the normalization steps (19)-(20) are repeated for all rows n where puncturing is required. The resulting $\mathbf{Q}_1$ is $\mathbf{W}$, and $\mathbf{L}_1$ is the desired punctured LTM $\mathbf{L}$.

In matrix form, we can write (17)-(18) using elementary matrices $\mathbf{E}_m =[\,e_{nj}]$, $1 \leq m \leq N$, which differ from $\mathbf{I}_N$ by a single elementary row operation, and defined as follows:

$$e_{nj} = \begin{cases} 1, & \text{if } j = n; \\ -l_{nm}/l_{mm}, & \text{if } j = m, j \in \mathcal{I}_n; \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

The product of these elementary matrices forms the unscaled matrices $\mathbf{L}_2 = (\mathbf{E}_n \cdots \mathbf{E}_1)\,\mathbf{L}_1$ and $\mathbf{Q}_2^* = \left(\mathbf{E}_n^* \cdots \mathbf{E}_1^*\right)\mathbf{Q}_1^*$.

The scaling operations (19)-(20) can be written using the diagonal matrix $\mathbf{D} = [d_n] \in \mathcal{R}^+$, where $d_n = 1\Big/\sqrt{\left[\mathbf{Q}_2^*\mathbf{Q}_2\right]_{nn}}$ and $[\,\cdot\,]_{nn}$ denotes the nth diagonal element. The desired (scaled) matrices are given by $\mathbf{W}^* = \mathbf{D}\mathbf{Q}_2^*$ and $\mathbf{L} = \mathbf{D}\mathbf{L}_2$.

For detection, the product $\mathbf{W}^*\mathbf{y}$ must be formed as well. This can simply be handled by first left-augmenting $\mathbf{y}$ to $\mathbf{H}$, and then performing QLD on the augmented matrix to form $\widetilde{\mathbf{Q}}\widetilde{\mathbf{L}} = \left[\mathbf{y}\,|\,\mathbf{H}\right]$. When carrying out the orthogonalization procedure, the same operations applied to the columns of $\mathbf{H}$ are applied to the augmented column. This results in $\widetilde{\mathbf{Q}} = [\mathbf{0}_{N\times 1}\,|\,\mathbf{Q}]$ and $\widetilde{\mathbf{L}} = \left[\widetilde{\mathbf{y}}\,|\,\mathbf{L}\right]$, where $\mathbf{QL} = \mathbf{H}$ and $\widetilde{\mathbf{y}} = \mathbf{Q}^*\mathbf{y}$, with $\widetilde{\mathbf{y}}$ essentially generated as a by-product of the decomposition. Next, when carrying out operations (17)-(18) followed (19)-(20) to puncture a given entry, these operations are also applied on the leftmost column of $\widetilde{\mathbf{L}}$ which contains $\widetilde{\mathbf{y}}$.

Algorithm 1 summarizes the steps of the proposed WLD scheme. Step 1 performs augmented QLD, while Step 2 performs nulling and normalization.

5.1 Complexity

We next analyze the complexity in terms of floating-point operations (flops) based on real multiplication (RMUL) and addition (RADD). We assume that real division and square-root operations are equivalent to a RMUL. Also, complex multiplication requires 4 RMUL and 2 RADD, while complex addition requires 2 RADD operations. Augmented QLD requires

$$\theta_1 = (4N^3 - N^2 - N)\mathsf{RADD} + (4N^3 + 3N^2)\mathsf{RMUL} \qquad (22)$$

flops, while puncturing requires

$$\theta_2 = \frac{2}{3}(8N^3 - 15N^2 + 4N - 12)\mathsf{RADD}\left(\frac{16}{3}N^3 - 7N^2 + \frac{8}{3}N - 20\right)\mathsf{RMUL} \qquad (23)$$

flops, assuming a $(1, (N-1)\times 1, N-1)$ structure. In comparison, [48] requires $(16N^4 - 4N^3)$RMUL and $(16N^4 - 4N^3)$RADD.

```
Algorithm 1: Optimized WL decomposition algorithm
$\mathbf{L} = \mathsf{WLD}(\mathbf{H}, \mathbf{y}, \mathcal{I})$
Input: Channel $\mathbf{H}$, $\mathbf{y}$, and puncturing sets $\mathcal{I} = \{\mathcal{I}_2, \cdots, \mathcal{I}_N\}$
Output: Transformed $\mathbf{L} = \left[\mathbf{W}^*\mathbf{y} \,|\, \mathbf{W}^*\mathbf{H}\right]$
# Note: col indices below account for augmented col in $\mathbf{Q}$ and $\mathbf{L}$
# Step 1: perform augmented QL decomposition
$\mathbf{Q} \leftarrow \left[\mathbf{y} \,|\, \mathbf{H}\right]$, $\mathbf{L} \leftarrow \mathbf{0}_{N\times(N+1)}$
for $n = N+1 : -1 : 2$ do    # loop over cols of $\mathbf{H}$
    $\mathbf{L}(n-1, n) \leftarrow \|\mathbf{q}_n\|$
    $\mathbf{q}_n \leftarrow \mathbf{q}_n / \mathbf{L}(n-1, n)$
    for $m = n-1 : -1 : 1$ do    # loop over cols left of $\mathbf{h}_n$
        $\mathbf{L}(n-1, m) \leftarrow \mathbf{q}_n^*\mathbf{q}_m$
        $\mathbf{q}_m \leftarrow \mathbf{q}_m - \mathbf{L}(n-1, m)\mathbf{q}_n$
    end
end
# Step 2: puncture using elementary row operations
for $n = 2 : N$ do    # loop over rows of $\mathbf{L}$ that need puncturing
    for $m = 1 : n-1, m \in \mathcal{I}_n$ do    # loop over cols to be punctured
        $\alpha \leftarrow \mathbf{L}(n, m+1) / \mathbf{L}(m, m+1)$
        $\mathbf{q}_{n+1} \leftarrow \mathbf{q}_{n+1} - \alpha^*\mathbf{q}_{m+1}$
        $\mathbf{L}(n, 1{:}m+1) \leftarrow \mathbf{L}(n, 1{:}m+1) - \alpha\mathbf{L}(m, 1{:}m+1)$
    end
    $\mathbf{L}(n, 1{:}n+1) \leftarrow \mathbf{L}(n, 1{:}n+1) / \|\mathbf{q}_{n+1}\|$
    $\mathbf{q}_{n+1} \leftarrow \mathbf{q}_{n+1} / \|\mathbf{q}_{n+1}\|$ # included only for completeness
end
```

6 Optimized detection algorithm

We next present a detection algorithm based on the proposed WLD scheme. We decompose $\mathbf{H}$ into T punctured LTMs having *identical* structure $(E, S\times E, S)$ where $T = N/E$ as follows. The N streams are decoupled E at a time in T steps by cyclically shifting the columns of $\mathbf{H}$ using the permutation

$$\pi_t(i) = \mathrm{mod}(i + (t-1)E - 1, N) + 1, \;\; i = 1, \cdots, N, \qquad (24)$$

for $t = 1, \cdots, T$. Each permuted $\mathbf{H}$ is then WL-decomposed into $\mathbf{W}^{(t)}, \mathbf{L}^{(t)}$, wherein the E leftmost columns of $\mathbf{L}^{(t)}$ correspond to E distinct decoupled streams. For example, to decouple all four streams of a 4×4 MIMO system using the $(2, 2\times 2, 2)$ structure of Figure 2b, $T = 2$ such decompositions are required, one on $[\,\mathbf{h}_1\;\mathbf{h}_2\;\mathbf{h}_3\;\mathbf{h}_4]$ to decouple streams 1 and 2 and one on $[\,\mathbf{h}_3\;\mathbf{h}_4\;\mathbf{h}_1\;\mathbf{h}_2]$ to decouple streams 3 and 4. To allow decoupled subsets to overlap, the permutation is altered to place a stream with low reliability in several detection sets. For example, to decompose the four streams in the previous example into overlapping $(2, 2\times 2, 2)$ structures, four decompositions are needed based on $[\,\mathbf{h}_1\;\mathbf{h}_2\;\mathbf{h}_3\;\mathbf{h}_4]$, $[\,\mathbf{h}_2\;\mathbf{h}_3\;\mathbf{h}_4\;\mathbf{h}_1]$, $[\,\mathbf{h}_3\;\mathbf{h}_4\;\mathbf{h}_1\;\mathbf{h}_2]$, and $[\,\mathbf{h}_4\;\mathbf{h}_3\;\mathbf{h}_2\;\mathbf{h}_1]$. For simplicity, we assume identical constellations $\mathcal{X}_n = \mathcal{O}$ on all layers.

One simple approach to low-complexity MIMO detection is zero-forcing. Left-multiplying $\mathbf{y}$ by $\mathbf{W}^{(t)*}$ to get $\mathbf{W}^{(t)*}\mathbf{y} = \widetilde{\mathbf{y}}^{(t)} = \mathbf{L}^{(t)}\mathbf{x} + \mathbf{W}^{(t)*}\mathbf{n}$ results in E decoupled streams corresponding to columns $\mathbf{h}_1, \cdots, \mathbf{h}_E$. These streams can be detected separately by slicing the layers individually to find their closest symbols.

A more powerful detection scheme takes full advantage of the punctured structure of $\mathbf{L}^{(t)}$. Instead of merely slicing to find the E closet symbols in the enumeration set, we enumerate all symbol vectors in the set $\mathcal{O}^E$ and then slice to find the closest symbols in the remaining S layers,

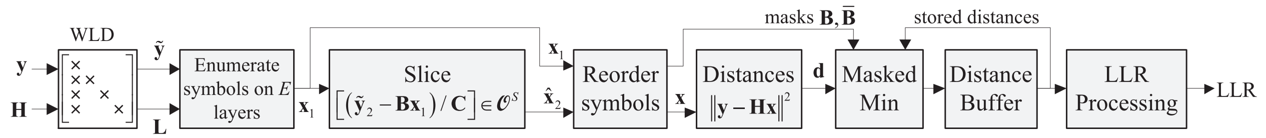
WLD
y
H
ỹ
L
Enumerate symbols on E layers
$\mathbf{x}_1$
Slice
$\left[\left(\tilde{\mathbf{y}}_2 - \mathbf{B}\mathbf{x}_1\right)/\mathbf{C}\right] \in \mathcal{O}^S$
$\hat{\mathbf{x}}_2$
Reorder symbols
x
Distances
$\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2$
d
masks $\mathbf{B}, \overline{\mathbf{B}}$
Masked Min
stored distances
Distance Buffer
LLR Processing
LLR

Figure 3 Block diagram of the proposed WL detection algorithm.

populating the Euclidean distance of the resulting symbol vectors along the way. First partition $\widetilde{\mathbf{y}}^{(t)}$, $\mathbf{L}^{(t)}$, and $\mathbf{x}$ as

$$\widetilde{\mathbf{y}}^{(t)} = \begin{bmatrix} \widetilde{\mathbf{y}}_1^{(t)} \\ \widetilde{\mathbf{y}}_2^{(t)} \end{bmatrix}, \quad \mathbf{L}^{(t)} = \begin{bmatrix} \mathbf{A}^{(t)} & \mathbf{0} \\ \mathbf{B}^{(t)} & \mathbf{C}^{(t)} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \tag{25}$$

where $\widetilde{\mathbf{y}}_1^{(t)} \in \mathcal{C}^{E\times 1}$, $\widetilde{\mathbf{y}}_2^{(t)} \in \mathcal{C}^{S\times 1}$, $\widetilde{\mathbf{A}}^{(t)} \in \mathcal{R}^{E\times E}$, $\widetilde{\mathbf{B}}^{(t)} \in \mathcal{R}^{S\times E}$, $\widetilde{\mathbf{C}}^{(t)} \in \mathcal{R}^{S\times S}$, $\mathbf{x}_1 \in \mathcal{O}^E$, and $\mathbf{x}_2 \in \mathcal{O}^S$. Then the symbol vector with minimum distance for the partitioned structure t is given by

$$\begin{aligned} \widetilde{\mathbf{x}}_t^{\mathrm{WL}} &\triangleq \arg\min_{\mathbf{x}\in\mathcal{X}} \left\| \widetilde{\mathbf{y}}^{(t)} - \mathbf{L}^{(t)}\mathbf{x} \right\|^2 \\ &= \arg\min_{\mathbf{x}_1\in\mathcal{O}^E} \left(\left\| \widetilde{\mathbf{y}}_1^{(t)} - \mathbf{A}^{(t)}\mathbf{x}_1 \right\|^2 + \left\| \widetilde{\mathbf{y}}_2^{(t)} - \mathbf{B}^{(t)}\mathbf{x}_1 - \mathbf{C}^{(t)}\widehat{\mathbf{x}}_2 \right\|^2 \right) \end{aligned} \tag{26}$$

$$\widehat{\mathbf{x}}_2 = \left\lfloor \left(\widetilde{\mathbf{y}}_2^{(t)} - \mathbf{B}^{(t)}\mathbf{x}_1 \right) / \mathbf{C}^{(t)} \right\rceil_{\mathcal{O}^S} \tag{27}$$

Note that the first argument of the slicing operator in (27) is a vector of length S since $\mathbf{C}^{(t)}$ is a diagonal matrix. Here slicing is applied to the individual elements of the vector over the constellation $\mathcal{O}$. The symbols in the vector $\widetilde{\mathbf{x}}_t^{\mathrm{WL}}$ are rearranged back to their normal order using (24), and the permuted vector is denoted as $\mathbf{x}_t^{\mathrm{WL}}$:

$$\mathbf{x}_t^{\mathrm{WL}} = \pi_t^{-1}\left(\widetilde{\mathbf{x}}_t^{\mathrm{WL}}\right).$$

The minimum distance d_t^{WL} itself is computed as the Euclidean distance of $\mathbf{y}$ from $\mathbf{H}\mathbf{x}_t^{\mathrm{WL}}$ (rather than the distance of $\widetilde{\mathbf{y}}^{(t)}$ from $\mathbf{L}^{(t)}\widetilde{\mathbf{x}}_t^{\mathrm{WL}}$):

$$d_t^{\mathrm{WL}} = \left\| \mathbf{y} - \mathbf{H}^{(t)}\mathbf{x}_t^{\mathrm{WL}} \right\|^2 \neq \left\| \widetilde{\mathbf{y}}^{(t)} - \mathbf{L}^{(t)}\widetilde{\mathbf{x}}_t^{\mathrm{WL}} \right\|^2. \tag{28}$$

To compute the LLRs, we expand distances similar to (26) when taking arg min to determine the symbol vector

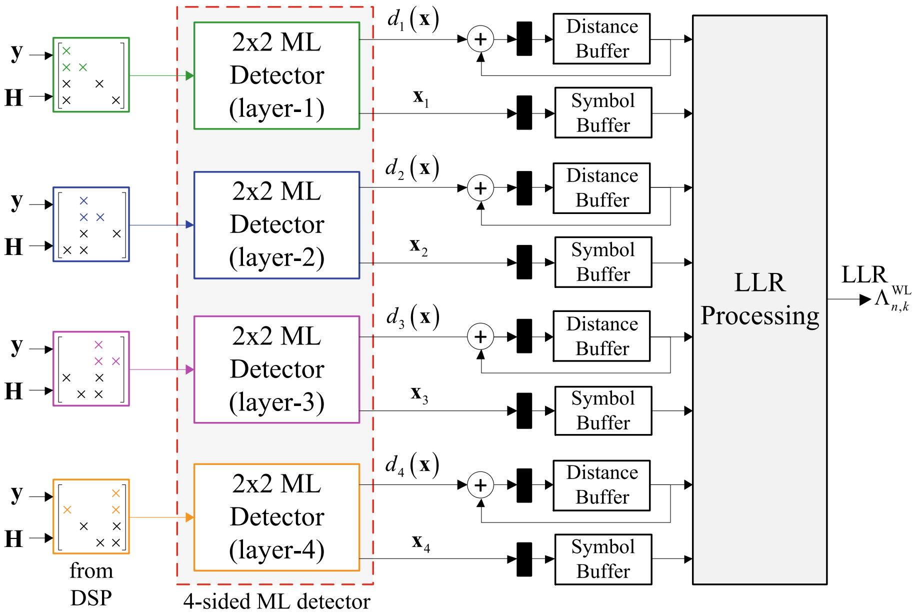

Figure 4 Block diagram of four-sided MAP detector.

$$\widetilde{\mathbf{u}}_{n,k,t}^{\mathrm{WL}} = \arg\min_{\mathbf{x}\in\mathcal{X}_{n,k}^{(0)}} \left\| \widetilde{\mathbf{y}}^{(t)} - \mathbf{L}^{(t)}\mathbf{x} \right\|^2 \qquad (29)$$

$$= \arg\min_{\mathbf{x}_1\in\mathcal{O}_{n,k}^{(0)}} \left(\left\| \widetilde{\mathbf{y}}_1^{(t)} - \mathbf{A}^{(t)}\mathbf{x}_1 \right\|^2 + \left\| \widetilde{\mathbf{y}}_2^{(t)} - \mathbf{B}^{(t)}\mathbf{x}_1 - \mathbf{C}^{(t)}\widehat{\mathbf{x}}_2 \right\|^2 \right)$$

where $\mathcal{O}_{n,k}^{(0)} = \left\{ \mathbf{x}_1 \in \mathcal{O}^E : b_{n,k} = 0 \right\}$, and $\widehat{\mathbf{x}}_2$ is defined as in (27) but with $\mathbf{x}_1 \in \mathcal{O}_{n,k}^{(0)}$. A similar computation is needed for the other hypothesis in the set $\mathcal{X}_{n,k}^{(1)}$ to determine:

$$\widetilde{\mathbf{v}}_{n,k,t}^{\mathrm{WL}} = \arg\min_{\mathbf{x}\in\mathcal{X}_{n,k}^{(1)}} \left\| \widetilde{\mathbf{y}}^{(t)} - \mathbf{L}^{(t)}\mathbf{x} \right\|^2 . \qquad (30)$$

Then (24) is applied to permute the symbols in $\widetilde{\mathbf{u}}_{n,k,t}^{\mathrm{WL}}$ and $\widetilde{\mathbf{v}}_{n,k,t}^{\mathrm{WL}}$ and form the reordered symbol vectors $\mathbf{u}_{n,k,t}^{\mathrm{WL}}$ and $\mathbf{v}_{n,k,t}^{\mathrm{WL}}$:

$$\mathbf{u}_{n,k,t}^{\mathrm{WL}} = \pi_t^{-1}\left(\widetilde{\mathbf{u}}_{n,k,t}^{\mathrm{WL}}\right),$$
$$\mathbf{v}_{n,k,t}^{\mathrm{WL}} = \pi_t^{-1}\left(\widetilde{\mathbf{v}}_{n,k,t}^{\mathrm{WL}}\right).$$

The LLR values are then computed as

$$\Lambda_{n,k,t}^{\mathrm{WL}} = \left\| \mathbf{y} - \mathbf{H}\mathbf{u}_{n,k,t}^{\mathrm{WL}} \right\|^2 - \left\| \mathbf{y} - \mathbf{H}\mathbf{v}_{n,k,t}^{\mathrm{WL}} \right\|^2 , \qquad (31)$$

for $n = 1, \cdots, N$, $k = 1 \cdots, \log_2 |\mathcal{O}|$, and $t = 1, \cdots, T$.

Finally, one simple way to approximate the ML distance and LLRs is by selecting the minimum over all t in (28) and (31), respectively:

$$d^{\mathrm{ML}} \approx \min_{t=1,\cdots,T}\left(d_t^{\mathrm{WL}}\right), \qquad \Lambda_{n,k}^{\mathrm{ML}} \approx \min_{t=1,\cdots,T}\left(\Lambda_{n,k,t}^{\mathrm{WL}}\right). \qquad (32)$$

Tighter LLRs can be produced by tracking global minimum distances rather than just minimizing over the per stream LLRs. Specifically, when using (26)-(27) for every stream t, $t = 1, \cdots, T$, instead of just retaining the minimum distance and its corresponding arg min symbol vector, a list of all $\left|\mathcal{O}^E\right|$ distances and their corresponding symbol vectors is populated. These T lists are then used to compute the LLR values by selecting the minimum distances for symbol vectors from these lists having 0 or 1 in the desired bit position where the LLR value is to be computed.

The pseudo-code in Algorithm 2 summarizes the steps of the proposed WLD detection algorithm. It produces tight LLRs according to the previous discussion, using the equations marked with $(\star)$. Figure 3 shows a block diagram describing the dataflow of the algorithm. For $N = 2$ layers, the algorithm is optimal and reduces to that in [45] since the $\mathbf{W}$s are unitary in this case. When $E = 1$ and distances in (28) are computed based on $\mathbf{L}$ instead of $\mathbf{H}$, the algorithm reduces to that in [48].

Algorithm 2: Optimized detection algorithm

```
$\Lambda$ = WDetector($\mathbf{H}, \mathbf{y}$)
Input: Channel matrix $\mathbf{H} \in \mathcal{C}^{N\times N}$; received vector
        $\mathbf{y} \in \mathcal{C}^{N\times 1}$
Output: LLR values $\Lambda \in \mathcal{R}^{q\times 1}$
Data: $T$ punctured matrices $\mathbf{L}$ of structure
        $(E, S\times E, S)$; puncturing set $\mathcal{I}$; distance
        vectors $\mathbf{d} =[d_j]$, $\mathbf{d}_0$, $\mathbf{d}_1 \in \mathcal{R}^{E\times 1}$; binary-mask
        matrix $\mathbf{B} \in \mathcal{F}_2^{E\times q}$
for $t = 1 : T$ do      # loop over all decompos. structures of $\mathbf{H}$
    $\pi_t(i) \leftarrow \mathrm{mod}(i + (t-1)E - 1, N) + 1$,  $i = 1, \cdots, N$
    $[\,\tilde{\mathbf{y}} \,|\, \mathbf{L}] \leftarrow$ WLD $(\mathbf{H}(:, \pi), \mathbf{y}, \mathcal{I})$  # WL decomposition function
    for $j = 1 : \left|\mathcal{O}^E\right|$ do      # loop over enum set E
        $\mathbf{x}_1 \leftarrow$ next partial symbol vector in$\mathcal{O}^E$
        $\widehat{\mathbf{x}}_2 \leftarrow \left\lfloor (\widetilde{\mathbf{y}}_2 - \mathbf{B}\mathbf{x}_1) / \mathbf{C} \right\rceil_{\mathcal{O}^S}$  # vector slice
        $\mathbf{x} = [x_i] \leftarrow [\mathbf{x}_1; \widehat{\mathbf{x}}_2] \in \mathcal{O}^N$  # for symbol vector
        $\mathbf{x} \leftarrow \left[x_{\pi(i)}\right]$  # reorder its symbols
        $d_j \leftarrow \left\|\mathbf{y} - \mathbf{H}\mathbf{x}\right\|^2$  # compute distances using $\mathbf{H}$
        $\mathbf{B}_{j,:} \leftarrow \mathrm{BinMask}(\mathbf{x})$  # get binary sequence; store in row j
    end
(⋆) $\mathbf{d}_1 \leftarrow \min(\mathbf{d}_1, \min(\mathbf{d}, \mathbf{B}))$  # masked-min operation using $\mathbf{B}$
(⋆) $\mathbf{d}_0 \leftarrow \min\left(\mathbf{d}_0, \min\left(\mathbf{d}, \overline{\mathbf{B}}\right)\right)$  # $\overline{\mathbf{B}}$ is binary complement of $\mathbf{B}$
end
$\Lambda \leftarrow \mathbf{d}_0 - \mathbf{d}_1$
```

6.1 Multi-core detector architectures

Depending on the target throughput and the number of antennas N in the MIMO systems, multiple 2 × 2 detector cores can be configured to construct an N-stream MIMO detector. Figure 4 shows a four-sided fully parallel 4 × 4 MIMO detector that uses four cores to process the four streams. Here distance buffering and accumulation are needed before LLR processing in order to adjust the individual LLRs according to earlier discussion. It is assumed that an external digital signal processor (DSP) must supply the WLD matrix inputs for all four streams according to the decompositions in (25). If chip area is the constraining factor, a MIMO detector can be built using a single core that is time-multiplexed among the four streams.

7 Performance simulation results

The coded bit-error rate (BER) performance of the proposed detection algorithm was evaluated through simulations of a MIMO system employing four transmit and four receive antennas. The channel encoder is based on the LTE turbo encoder specification [52] with interleaver length 1,024, using 16-QAM and 64-QAM modulation constellations. The channel entries are assumed to be independent and identically distributed (i.i.d.) complex Gaussian random variables. At the receiver end, we assume perfect channel knowledge. The turbo decoder implements the true *A Posteriori* Probability algorithm, and performs four (full) decoding iterations.

Figure 5 compares the BER versus SNR per receive antenna of the proposed WLD scheme with $E = 1$ and 2 structures, versus ML, ZF, the approach of [48], and the sphere decoder with radius clipping [30], for 16-QAM. Both overlapping and non-overlapping subsets are considered. Furthermore, two scenarios for distance computations in (28) are followed; one based on **H** and one on **L**. The curve labeled WLD-H1 corresponds to the proposed WLD detection scheme with $E = 1$ and distance computations based on **H** in (28), while the curve labeled WLD-L1 corresponds to the proposed scheme with $E = 1$ but with distance computations based on **L** in (28). Similarly for the curves labeled WLD-H2 and WLD-L2, but with $E = 2$ and non-overlapping decomposition sets. Finally the curve labeled WLD-H2ov corresponds to $E = 2$ but with four overlapping $(2, 2 \times 2, 2)$ structures.

As shown in the figure, the plots demonstrate that the proposed WLD algorithm with $E = 2$ using **H** distances with overlapping subsets performs virtually as ML and is less than 0.1 dB away from ML with no overlapping subsets. Furthermore, the WLD-L2 scheme performs much worse than WLD-H2. For decompositions with single streams, **L** distances perform better than **H** distances; the WLD-H1 scheme exhibits an error floor. Compared to the well-known sphere decoding algorithm with radius-clipping [30], the proposed WLD-based schemes achieve better performance especially with $E = 2$ structures.

Figure 6 compares the BER performance of the WLD schemes using 64-QAM constellations. The plots demonstrate again that the WLD scheme with $E = 2$ using **H** distances and overlapping subsets performs very close to ML. It is worth mentioning that WLD-H2 again performs very close to WLD-H2ov (and hence to ML), which has significant hardware saving implications. Note here that the WLD-L2ov scheme using **L**-distances with $E = 2$ overlapping decompositions performs the worst among the WLD schemes.

Finally, another important advantage of the proposed schemes is the significant reduction in simulation time observed when generating the BER plots. On a typical multi-core server machine, one simulation point takes in the order of few hours to complete, while the sphere decoder and ML approaches require few days. This is very valuable for designers when doing rapid evaluations and tradeoff analyses of various MIMO system features.

8 Conclusions

A low-complexity MIMO subspace detection algorithm has been presented. By decomposing the channel matrix of an M-stream MIMO system into a generalized elementary matrix structure, the detection problem becomes a generalization of that of a two-stream detection problem, which admits a simple architecture suitable for high-speed implementation. Multiple two-stream detector cores can be employed in parallel to improve detection throughput. The channel decomposition scheme is cast in terms of the standard Gram-Schmidt QL decomposition, which is supported in most modern DSPs.

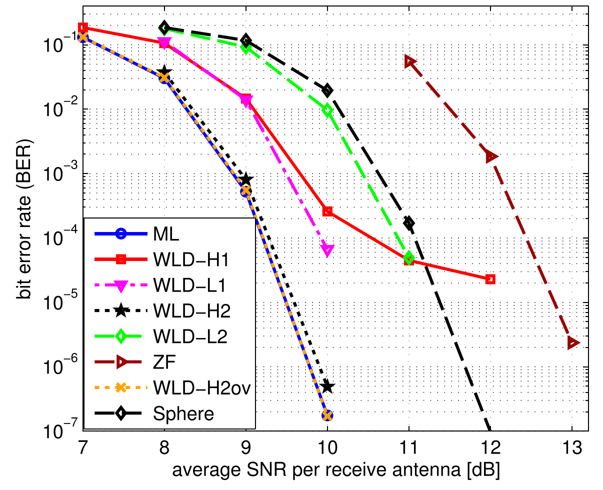

Figure 5 **BER vs. SNR plots for 16-QAM constellation.**

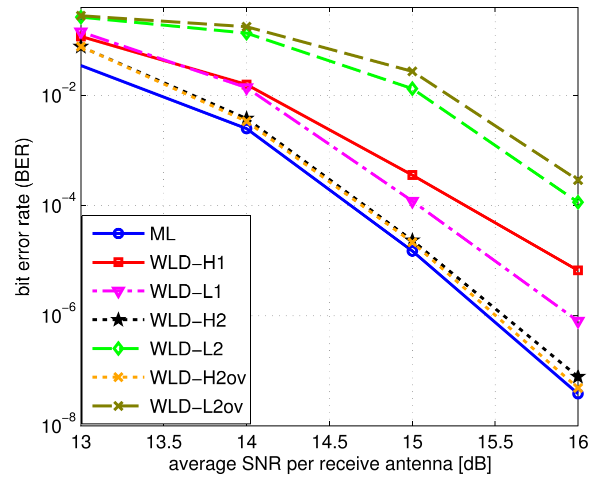

Figure 6 **BER vs. SNR plots for 64-QAM constellation.**

Competing interests

The author declares that he has no competing interests.

Received: 24 October 2014 Accepted: 9 February 2015
Published online: 28 March 2015

References

1. A Paulraj, R Nabar, D Gore, *Introduction to Space-Time Wireless Communications*. (Cambridge Univ. Press, Cambridge, U.K, 2003)
2. E Viterbo, J Boutros, A universal lattice code decoder for fading channels. IEEE Trans. Inf. Theory. **45**(5), 1639–1642 (1999)
3. O Damen, A Chkeif, J-C Belfiore, Lattice code decoder for space-time codes. IEEE Commun. Lett. **4**(5), 161–163 (2000)
4. E Agrell, T Eriksson, A Vardy, K Zeger, Closest point search in lattices. IEEE Trans. Inf. Theory. **48**(8), 2201–2214 (2002)
5. GJ Foschini, Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. Bell Labs Tech. J. **1**(2), 41–59 (1996)
6. Y Jiang, J Li, WW Hager, Joint transceiver design for MIMO communications using geometric mean decomposition. IEEE Trans. Signal Process. **53**(10), 3791–3803 (2005)
7. Y Jiang, J Li, WW Hager, Uniform channel decomposition for MIMO communications. IEEE Trans. Signal Process. **53**(11), 4283–4294 (2005)
8. SL Ariyavisitakul, J Zheng, E Ojard, J Kim, Subspace beamforming for near-capacity MIMO performance. IEEE Trans. Signal Process. **56**(11), 5729–5733 (2008)
9. Y Chen, ST Brink, in *Proc. IEEE Int. Symp. Personal Indoor and Mobile Radio Commun. (PIMRC)*. Near-capacity MIMO subspace detection (Toronto, Canada, 2011), pp. 1733–1737
10. E Biglieri, R Calderbank, A Constantinides, A Goldsmith, A Paulraj, HV Poor, *MIMO Wireless Communications*. (Cambridge Univ. Press, Cambridge, U.K, 2007)
11. B Hassibi, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal, Process. (ICASSP)*. An efficient square-root algorithm for BLAST (Istanbul, Turkey, 2000), pp. 5–9
12. GD Golden, JG Foschini, RA Valenzuela, PW Wolniansky, Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture. IEE Electron. Lett. **35**(1), 14–15 (1999)
13. D Wübben, R Böhnke, V Kühn, K Kammeyer, in *Proc. IEEE Vehicular Technol. Conf. (VTC)*. MMSE extension of V-BLAST based on sorted QR decomposition (Orlando, Florida, 2003), pp. 508–512
14. C Studer, S Fateh, D Seethaler, ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation. IEEE Trans. Syst. Sci. Cybern. **47**(7), 1754–1765 (2011)
15. E Viterbo, E Biglieri, in *14ème Colloque GRETSI*. A universal decoding algorithm for lattice codes (Juan-Les-Pins, France, 1993), pp. 611–614
16. B Hochwald, S ten Brink, Achieving near-capacity on a multiple-antenna channel. IEEE Trans. Commun. **51**(3), 389–399 (2003)
17. B Hassibi, H Vikalo, On sphere decoding algorithm. I. Expected complexity. IEEE Trans. Signal Process. **53**(8), 2806–2818 (2005)
18. J Jaldén, B Ottersten, On the complexity of sphere decoding in digital communications. IEEE Trans. Signal Process. **53**(4), 1474–1484 (2005)
19. D Seethaler, J Jaldén, C Studer, H Bölcskei, On the complexity distribution of sphere decoding. IEEE Trans. Inf. Theory. **57**(9), 5754–5768 (2011)
20. K-W Wong, C-Y Tsui, RS-K Cheng, W-H Mow, in *Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS)*. A VLSI architecture of a *K*-best lattice decoding algorithm for MIMO channels, vol. 3 (Scottsdale, Arizona, 2002), pp. 273–276
21. M Wenk, M Zellweger, A Burg, N Felber, W Fichtner, in *Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS)*. K-best MIMO detection VLSI architectures achieving up to 424Mbps (Island of Kos, Greece, 2006), pp. 1151–1154
22. S Mondal, A Eltawil, C-A Shen, K Salama, Design and implementation of a sort free *K*-best sphere decoder. IEEE Trans. VLSI Syst. **18**(10), 1497–1501 (2010)
23. L Liu, F Ye, X Ma, T Zhang, J Ren, A 1.1-Gb/s 115-pJ/bit configurable MIMO detector using 0.13 μm CMOS technology. IEEE Trans. Circuits Syst. II. **57**(9), 701–705 (2010)
24. C-A Shen, A Eltawil, K Salama, Evaluation framework for *K*-best sphere decoders. J. Circuits Syst. Comput. **19**(5), 975–995 (2010)
25. M Shabany, P Gulak, A 675 Mbps, 4×4 64-QAM K-Best MIMO detector in 0.13 μm CMOS. IEEE Trans. VLSI Syst. **20**(1), 135–147 (2012)
26. M Mahdavi, M Shabany, Novel MIMO detection algorithm for high-order constellations in the complex domain. IEEE Trans. VLSI Syst. **21**(5), 834–847 (2013)
27. D Garrett, L Davis, S ten Brink, B Hochwald, G Knagge, Silicon complexity for maximum likelihood MIMO detection using spherical decoding. IEEE J. Solid-State Circuits. **39**(9), 1544–1552 (2004)
28. Z Guo, P Nilsson, in *Proc. IEEE CAS Symp. Emerging Technologies*. A VLSI architecture of the Schnorr-Euchner decoder for MIMO systems, vol. 1 (Shanghai, China, 2004), pp. 65–68
29. A Burg, M Borgmann, M Wenk, M Zellweger, W Fichtner, H Bölcskei, VLSI implementation of MIMO detection using the sphere decoding algorithm. IEEE J. Solid-State Circuits. **40**(7), 1566–1577 (2005)
30. C Studer, A Burg, H Bölcskei, Soft-output sphere decoder: algorithms and VLSI implementation. IEEE J. Sel. Areas Commun. **26**(2), 290–300 (2008)
31. C-H Yang, D Markovic, A flexible DSP architecture for MIMO sphere decoding. IEEE Trans. Circuits Syst. I. **56**(10), 2301–2314 (2009)
32. C-H Yang, D Markovic, in *European Solid-State Circuits Conf. (ESSCIRC)*. A 2.89 mW 50 GOPS 16×16 16-core MIMO sphere decoder in 90 nm CMOS (Athens, Greece, 2009), pp. 344–347
33. F Borlenghi, EM Witte, G Ascheid, H Meyr, A Burg, in *IEEE Asian Solid State Circutis Conf. (A-SSCC)*. A 772 Mbit/s 8.81 bit/nJ 90 nm CMOS soft-input soft-output sphere decoder (Jeju, Korea, 2011), pp. 297–300
34. L Liu, J Lofgren, P Nilsson, Area-efficient configurable high-throughput signal detector supporting multiple MIMO modes. IEEE Trans. Circuits Syst. I. **59**(9), 2085–2096 (2012)
35. Y Sun, JR Cavallaro, Trellis-search based soft-input soft-output MIMO detector: algorithm and VLSI architecture. IEEE Trans. Signal Process. **60**(5), 2617–2627 (2012)
36. X Chen, G He, J Ma, VLSI implementation of a high-throughput iterative fixed-complexity sphere decoder. IEEE Trans. Circuits Syst. II. **60**(5), 272–276 (2013)
37. MM Mansour, S Alex, MA Jalloul, Reduced complexity soft-output MIMO sphere detectors – Part I Algorithmic optimizations. IEEE Trans. Signal Process. **62**(21), 5505–5520 (2014)
38. MM Mansour, S Alex, MA Jalloul, Reduced complexity soft-output MIMO sphere detectors – Part II Architecutral optimizations. IEEE Trans. Signal Process. **62**(21), 5521–5535 (2014)
39. M-Y Huang, P-Y Tsai, Toward multi-gigabit wireless: design of high-throughput MIMO detectors with hardware-efficient architecture. IEEE Trans. Circuits Syst. I. **61**(2), 613–624 (2014)
40. J Choi, A Singer, J Lee, N-I Cho, Improved linear soft-input soft-output detection via soft feedback successive interference cancellation. IEEE Trans. Commun. **58**(3), 986–996 (2010)
41. RC de Lamare, R Sampaio-Neto, Adaptive reduced-rank equalization algorithms based on alternating optimization design techniques for MIMO systems. IEEE Trans. Veh. Technol. **60**(6), 2482–2494 (2011)
42. T Hwang, Y Kim, H Park, Energy spreading transform approach to achieve full diversity and full rate for MIMO systems. IEEE Trans. Signal Process. **60**(12), 6547–6560 (2012)
43. D Persson, J Kron, M Skoglund, EG Larsson, Joint source-channel coding for the MIMO broadcast channel. IEEE Trans. Signal Process. **60**(4), 2085–2090 (2012)
44. RC de Lamare, Adaptive and iterative multi-branch MMSE decision feedback detection algorithms for multi-antenna systems. IEEE Trans. Wireless Commun. **12**(10), 5294–5308 (2013)
45. M Siti, MP Fitz, in *Proc. IEEE Int. Conf. Commun. (ICC)*. A novel soft-output layered orthogonal lattice detector for multiple antenna communications, vol. 4 (Istanbul, Turkey, 2006), pp. 1686–1691
46. M Siti, MP Fitz, in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*. On layer ordering techniques for near-optimal MIMO detectors (Hong Kong, 2007), pp. 1199–1204
47. MS Yee, in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal, Process. (ICASSP)*. Max-log-MAP sphere decoder, vol. 3 (Philadelphia, PA, 2005), pp. 1013–1016
48. E Ojard, S Ariyavisitakul, Method and system for approximate maximum likelihood (ML) detection in a multiple input multiple output (MIMO) receiver. U. S. Patent No. 20090074114 (2009). http://www.google.com/patents/US20090074114

49. GH Golub, CFV Loan, *Matrix Computations*, 3rd edn. (Johns Hopkins Univ. Press, Baltimore, MD, 1996)
50. RC-H Chang, C-H Lin, K-H Lin, C-L Huang, F-C Chen, Iterative QR decomposition architecture using the modified Gram-Schmidt algorithm for MIMO systems. IEEE Trans. Circuits Syst. I. **57**(5), 1095–1102 (2010)
51. D Wübben, R Böhnke, J Rinas, V Kühn, K Kammeyer, Efficient algorithm for decoding layered space-time codes. IEE Electron. Lett. **37**(22), 1348–1350 (2001)
52. Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation. TS 36.211 3GPP. http://www.3gpp.org