A Near-ML MIMO Subspace Detection Algorithm

Mohammad M. Mansour, *Senior Member, IEEE*

***Abstract*—A low-complexity MIMO detection scheme is presented that decomposes a MIMO channel into multiple decoupled subsets of streams that can be detected separately. The scheme employs QL decomposition followed by elementary matrix operations to transform the channel matrix into a generalized elementary structure matching the subsets of streams to be detected. The proposed scheme avoids matrix inversion operations, and allows subsets to overlap thus achieving better diversity gain. Simulations demonstrate that this approach performs to within a few tenths of a dB from the optimum detection algorithm.**

***Index Terms*—Log-likelihood ratios (LLRs), maximum likelihood (ML), MIMO detection, subspace detection.**

I. Introduction

MULTIPLE-INPUT MULTIPLE-OUTPUT (MIMO) antenna systems have become mainstream technology for achieving high spectral efficiency in modern wireless communications standards. Detection of spatially multiplexed MIMO streams plays a key role in receiver design in terms of performance and complexity [1], and has remained an active area of research. MIMO detection schemes are either based on joint-stream detection to separate the streams, such as 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 host of MIMO detectors have appeared in the literature, offering various performance-complexity tradeoffs. Suboptimal zero-forcing and MMSE detectors, as well as the nonlinear parallel and SIC schemes, require relatively low complexity but sacrifice performance. Sphere detectors require substantially higher complexity, but achieve ML performance [2]–[4], [10].

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 $\mathbf{H}$ is uniformly decomposed into identical parallel subchannels using geometric mean decomposition (GMD). In [8], a related scheme was presented in which $\mathbf{H}$ is block-wise diagonalized to reduce 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 to overlap, additional diversity can be gathered by putting a low reliable data stream into several detection sets.

The LORD algorithm [11], [12] can be viewed as a special class of subspace MIMO detectors. It achieves log-max [13] optimal performance on 2-transmit antennas, but its performance degrades for larger number of antennas. In [14], the LORD algorithm was generalized to 4-transmit antennas by using matrix inversion to decompose $\mathbf{H}$ into single streams.

Contributions: In this letter, we propose an efficient MIMO detector that jointly detects subsets of decoupled streams by transforming $\mathbf{H}$ into a generalized elementary matrix. This is achieved by applying QL decomposition followed by elementary matrix operations on $\mathbf{H}$, thus avoiding the need for matrix inversion. The streams in the subsets can be detected in parallel, and are allowed to overlap, resulting in additional diversity gain. For 2 streams, the algorithm is optimal and reduces to the LORD algorithm. When the subsets include single streams only, the algorithm reduces to that of [14].

The rest of the letter is organized as follows. Section II introduces the system model, and Section III reviews ML detection for 2 streams. Sections IV, V present the proposed matrix decomposition scheme and a simplified construction using QLD. The detection algorithm is detailed in Section IV. Section VII presents simulation results.

II. System Model

Consider a MIMO system with N transmit antennas and $M \geq N$ receive antennas. Assuming perfect channel knowledge at the receiver, the equivalent complex baseband input-output system relation is $\mathbf{y} = \mathbf{Hx} + \mathbf{n}$, 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 $\mathbb{E}[x_n^* x_n] = 1$, and is formed from a set of q_n coded bit-interleaved sequence $\mathbf{b}_n = (b_{n,1}, b_{n,2}, \cdots, b_{n,q_n})$ over the binary field $\mathcal{F}_2$. 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 $\mathbb{E}[\mathbf{nn}^*] = \sigma_{\mathbf{n}}^2 \mathbf{I}_M$. 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.

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

$$d(\mathbf{x}) \triangleq \|\mathbf{y} - \mathbf{Hx}\|^2 = \|\tilde{\mathbf{y}} - \mathbf{Lx}\|^2, \tag{1}$$

where $\mathbf{H} = \mathbf{QL}$ is the QL decomposition (QLD) 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 $\tilde{\mathbf{y}} = \mathbf{Q}^*\mathbf{y} = \mathbf{Lx} + \mathbf{Q}^*\mathbf{n} \in \mathcal{C}^{N\times 1}$. Since $\mathbf{Q}$ is unitary, it preserves Euclidean norm and noise statistics. For equiprobable symbols, a "hard-decision" ML MIMO detector finds $\mathbf{x}^{\mathrm{ML}} \in \mathcal{X}$ such that $\mathbf{Hx}^{\mathrm{ML}}$

Manuscript received August 07, 2014; revised September 09, 2014; accepted September 09, 2014. Date of publication September 16, 2014; date of current version October 08, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rodrigo C. de Lamare.

The author is with Department of Electrical and Computer Engineering, American University of Beirut, Beirut, Lebanon (e-mail: mmansour@ieee.org).

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

Digital Object Identifier 10.1109/LSP.2014.2357991

is closest to $\mathbf{y}$ in $\mathcal{C}^{M\times 1}$ (or $\mathbf{L}\mathbf{x}^{\mathrm{ML}}$ is closest to $\tilde{\mathbf{y}}$ in $\mathcal{C}^{N\times 1}$). This is an integer least-squares problem [4] of the form

$$d^{\mathrm{ML}} = \min_{\mathbf{x}\in\mathcal{X}} \|\tilde{\mathbf{y}} - \mathbf{L}\mathbf{x}\|^2, \mathbf{x}^{\mathrm{ML}} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\tilde{\mathbf{y}} - \mathbf{L}\mathbf{x}\|^2. \quad (2)$$

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

$$\Lambda^{\mathrm{ML}}_{n,k} = \min_{\mathbf{x}\in\mathcal{X}^{(0)}_{n,k}} d(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}^{(1)}_{n,k}} d(\mathbf{x}), \quad (3)$$

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

III. Optimum 2×2 Soft-Output MIMO Detection

The ML solution requires computing $\prod_{n=1}^{N} Q_n$ distance metrics. For $N = 2$, a simplification [11] can be applied to reduce the number of computations from Q_1Q_2 to $Q_1 + Q_2$ by triangularizing $\mathbf{H}$ as $\mathbf{H} = \mathbf{Q}_1\mathbf{L}_1$ with $\mathbf{Q}_1$ being unitary:

$$\mathbf{y} - \mathbf{H}\mathbf{x} \rightarrow \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{bmatrix} - \begin{bmatrix} a_1 & 0 \\ c_1 & b_1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \tilde{\mathbf{y}}_1 - \mathbf{L}_1\mathbf{x} \quad (4)$$

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

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}} \|\tilde{\mathbf{y}}_1 - \mathbf{L}_1\mathbf{x}\|^2 &= \min_{\substack{x_1\in\mathcal{X}_1\\ x_2\in\mathcal{X}_2}} \left(|\tilde{y}_1 - a_1x_1|^2 + |\tilde{y}_2 - c_1x_1 - b_1x_2|^2\right)\\ &= \min_{x_1\in\mathcal{X}_1} \left(|\tilde{y}_1 - a_1x_1|^2 + |\tilde{y}_2 - c_1x_1 - b_1\hat{x}_2|^2\right)\\ &\triangleq \min_{x_1\in\mathcal{X}_1} d_1(x_1) \end{aligned} \quad (5)$$

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

$$\hat{x}_2 = \lfloor(\tilde{y}_2 - c_1x_1)/b_1\rceil_{x_2} \in \mathcal{X}_2. \quad (6)$$

Hence (5) requires only $|\mathcal{X}_1| = Q_1$ distance computations. The LLRs of the bits in symbol x_1 can simply be obtained as

$$\Lambda^{\mathrm{ML}}_{1,k} = \min_{x_1\in\mathcal{X}^{(0)}_{1,k}} d_1(x_1) - \min_{x_1\in\mathcal{X}^{(1)}_{1,k}} d_1(x_1), k = 1,\cdots,q_1. \quad (7)$$

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:

$$\mathbf{y} - \mathbf{H}\mathbf{x} \rightarrow \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} - \begin{bmatrix} 0 & a_2 \\ b_2 & c_2 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \bar{\mathbf{y}}_2 - \mathbf{L}_2\mathbf{x} \quad (8)$$

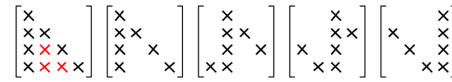
where now $\bar{\mathbf{y}}_2 = \mathbf{Q}_2^*\mathbf{y}$, and $a_2, b_2 \in \mathcal{R}^+$ and $c_2 \in \mathcal{C}$. Then

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}} \|\bar{\mathbf{y}}_2 - \mathbf{L}_2\mathbf{x}\|^2 &= \min_{x_2\in\mathcal{X}_2} \left(|\bar{y}_1 - a_2x_2|^2 + |\bar{y}_2 - c_2x_2 - b_2\hat{x}_1|^2\right)\\ &\triangleq \min_{x_2\in\mathcal{X}_2} d_2(x_2) \end{aligned} \quad (9)$$

where $\hat{x}_1 = \lfloor(\bar{y}_2 - c_2x_2)/b_2\rceil_{\mathcal{X}_1}$. The bit LLRs for x_2 are

$$\Lambda^{\mathrm{ML}}_{2,k} = \min_{x_2\in\mathcal{X}^{(0)}_{2,k}} d_2(x_2) - \min_{x_2\in\mathcal{X}^{(1)}_{2,k}} d_2(x_2), \quad k = 1,\cdots,q_2. \quad (10)$$

Since $\mathbf{Q}_1, \mathbf{Q}_2$ are unitary, the ML solutions in (5), (9) are identical. To find the hard-decision ML solution, only 1-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 2, and the minimum is selected. However, to generate soft LLRs, 2-sided decompositions are needed, and 2 lists of distances for $n = 1$ *and* 2 must be computed to select the appropriate minima in (7) and (10).



(a) (b) (c) (d) (e)

Fig. 1. 4×4 structures: (a) Full; (b)-(e) punctured structures for every layer.

IV. Extensions to Higher-Order Layers

The previous optimization cannot be extended to $N \geq 3$ layers because $\mathbf{L}$ includes off-diagonal terms that prevent computing the ML solution by enumerating symbols on one layer and finding the minima through slicing individually on all other layers in parallel. In Fig. 1(a), the red-marked entries in $\mathbf{L}$ imply that the ML solution requires enumerating symbols on $N - 1$ layers and slicing only on the last layer, as is done in tree-based detectors, hence still requiring $O(\prod_n Q_n)$ complexity.

A desirable structure of $\mathbf{H}$ for a 4-layer MIMO system is shown in Fig. 1(b), where the red-marked entries are zeroed-out. By enumerating symbols on layer 1, the minimum distances on layers 2 to 4 can be searched for in parallel through slicing only on the corresponding layers, similar to the 2-layer system. This suffices to compute the LLRs associated with the layer-1 bits. A similar process is repeated by decomposing $\mathbf{H}$ according to Figs. 1(c)–(e) [14] to compute the LLRs for bits associated with layers 2 to 4.

Fig. 2 shows other possible 4×4 "punctured" structures. They differ in 1) the number of layers over which symbols are enumerated (*enumeration set* of size S), 2) the submatrix structure used to propagate these enumerated symbols and cancel their interference effect from other layers (*interference cancelation set* of size E), and 3) the number of layers in which the minimum distance and associated symbol can be obtained by slicing after interference cancelation (*slicer set* of size $S\times E$). We refer to this structure using the triplet $(E, S\times E, S)$. 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.

A. 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 $N = M$. 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 $\mathbf{W}^*\mathbf{H} = \mathbf{L}$. 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 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 an 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

$$\mathrm{diag}\,(\mathbf{W}^*\mathbf{W}) = [11\cdots 1]^T_{1\times N}. \quad (11)$$

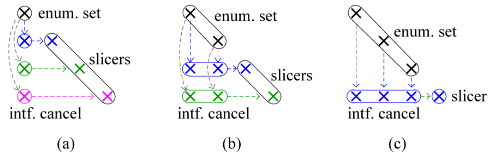


Fig. 2. (a) $(1, 3\times 1, 3)$, (b) $(2, 2\times 2, 2)$, (c) $(3, 1\times 3, 1)$ punctured structure.

This guarantees that the transformed noise vector $\mathbf{W}^*\mathbf{n}$ has an unaltered covariance matrix $\mathbb{E}[\mathbf{W}^*\mathbf{n}\mathbf{n}^*\mathbf{W}] = \sigma_{\mathbf{n}}^2\mathbf{I}_N$.

B. Proposed WL Decomposition (WLD) 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 $n \in \mathcal{I}$. 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 $\widetilde{\mathbf{w}}_n = \mathbf{P}_{\mathcal{I}_n}^\perp\mathbf{h}_n$, where

$$\mathbf{P}_{\mathcal{I}_n}^\perp = \mathbf{I}_N - \mathbf{H}_{\mathcal{I}_n}(\mathbf{H}_{\mathcal{I}_n}^*\mathbf{H}_{\mathcal{I}_n})^{-1}\mathbf{H}_{\mathcal{I}_n}^* \tag{12}$$

and $\mathbf{H}_{\mathcal{I}_n} = \{\mathbf{h}_m | m \in \mathcal{I}_n\}$. To satisfy (11), first note that

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

where the second equality follows since $\mathbf{P}_{\mathcal{I}_n}^\perp$ is a projection matrix, and hence $(\mathbf{P}_{\mathcal{I}_n}^\perp)^* = \mathbf{P}_{\mathcal{I}_n}^\perp$ and $\mathbf{P}_{\mathcal{I}_n}^\perp\mathbf{P}_{\mathcal{I}_n}^\perp = \mathbf{P}_{\mathcal{I}_n}^\perp$. Hence the normalized vector $\mathbf{w}_n = \widetilde{\mathbf{w}}_n/\|\widetilde{\mathbf{w}}_n\|$ with $\|\widetilde{\mathbf{w}}_n\| = \sqrt{\mathbf{h}_n^*\mathbf{P}_{\mathcal{I}_n}^\perp\mathbf{h}_n}$, has unit length. We have the following result:

Theorem 1: 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 $\mathbf{P}_{\mathcal{I}_n}^\perp$. Let $\mathbf{D} = [d_n] \in \mathcal{R}^+$ be a diagonal matrix whose entries are given by $d_n = 1/\sqrt{\mathbf{h}_n^*\mathbf{P}_{\mathcal{I}_n}^\perp\mathbf{h}_n}$, $n = 1, \cdots, N$. Then the matrix

$$\mathbf{W}^* = \mathbf{D}\begin{bmatrix} \mathbf{h}_1^*\mathbf{P}_{\mathcal{I}_1}^\perp \\ \mathbf{h}_2^*\mathbf{P}_{\mathcal{I}_2}^\perp \\ \vdots \\ \mathbf{h}_N^*\mathbf{P}_{\mathcal{I}_N}^\perp \end{bmatrix}, \tag{13}$$

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 (11).

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 $m \notin \mathcal{I}_n, 1 \le m \le N$. Hence $\mathbf{w}_n$ nulls the nth row of $\mathbf{H}$ only at the column indices given in $\mathcal{I}_n$. ■

For example, in a 4×4 MIMO system, choosing the puncturing sets as $\mathcal{I}_1 = \{2, 3, 4\}, \mathcal{I}_2 = \{3, 4\}, \mathcal{I}_3 = \{2, 4\}, \mathcal{I}_4 = \{2, 3\}$, results in a LTM $\mathbf{L}$ with entries $l_{32} = l_{42} = l_{43} = 0$ as (Fig. 2(a)):

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

The $\mathbf{W}, \mathbf{L}$ pair 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: 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

$$\mathbf{L}_1 = \mathbf{R}_1^*(\mathbf{R}_2^*)^{-1}\mathbf{L}_2, \quad \mathbf{W}_1 = \mathbf{R}_1\mathbf{R}_2^{-1}\mathbf{W}_2 \tag{14}$$

The proof is omitted due to lack of space.

V. Reduced-Complexity WL Decomposition

A brute force approach for computing $\mathbf{W}$ involves extensive matrix inversion, which is computationally expensive and prone to numerical error when implemented with finite precision. We develop an efficient scheme to determine $\mathbf{W}$ using the modified Gram-Schmidt orthogonalization procedure [15] and elementary matrix operations. From lemma 1, any other $\mathbf{W}$ that produces an identical structure is related by (14).

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 $1 < n \le N$ 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 $(\mathbf{q}_n^* - \mathbf{q}_m^*\frac{l_{nm}}{l_{mm}})\mathbf{h}_m = 0$. Therefore

$$\mathbf{q}_n = \mathbf{q}_n - \mathbf{q}_m l_{nm}^*/l_{mm} \tag{15}$$

$$l_{nj} = l_{nj} - l_{mj}l_{nm}/l_{mm}, \text{for } j = 1, \cdots, m, \tag{16}$$

puncture the required entry and update $\mathbf{Q}_1$ accordingly. These operations are repeated for all other 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:

$$l_{nj} = l_{nj}/\left\|\mathbf{q}_n\right\|, \text{for } j = 1, \cdots, n \tag{17}$$

$$\mathbf{q}_n = \mathbf{q}_n/\left\|\mathbf{q}_n\right\| \tag{18}$$

The operations in (15)–(16) followed by the normalization steps (17)–(18) 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 $\mathbf{L}$.

In matrix form, we can write (15)–(16) using elementary matrices $\mathbf{E}_m = [e_{nj}], 1 \le m \le N$, which differ from $\mathbf{I}_N$ by a single elementary row operation, and defined as follows:

$$e_{nj} = \begin{cases} 1, & \text{if } j = n; \\ -l_{nm}/l_{mm}, & \text{if } j = m, j \in \mathcal{I}_n; \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

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

The scaling operations (17)–(18) can be written using the diagonal matrix $\mathbf{D} = [d_n] \in \mathcal{R}^+$, where $d_n = 1/\sqrt{[\mathbf{Q}_2^*\mathbf{Q}_2]_{nn}}$ 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 is handled by first left-augmenting $\mathbf{y}$ to $\mathbf{H}$, and then performing QLD on the augmented matrix to form $\widetilde{\mathbf{Q}}\widetilde{\mathbf{L}} = [\mathbf{y}|\mathbf{H}]$. 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 $\widetilde{\mathbf{L}} = [\widetilde{\mathbf{y}}|\mathbf{L}]$, where $\mathbf{Q}\mathbf{L} = \mathbf{H}$ and $\widetilde{\mathbf{y}} = \mathbf{Q}^*\mathbf{y}$, with $\widetilde{\mathbf{y}}$ essentially generated as a by-product.. Next, when carrying out operations (15)–(16) followed by (17)–(18) to puncture a given entry, these operations are also applied on the leftmost column of $\widetilde{\mathbf{L}}$.

We next analyze the complexity in terms of floating-point operations (flops) based on real multiplication (RMUL) and addition (RADD). Assume that real division and square-root opertions are equivalent to a RMUL. Also, complex multiplication requires 4 RMUL and 2 RADD, while complex addition requires 2 RADD. Augmented QLD requires $\theta_1 = (4N^3 - N^2 - N)\,\text{RADD} + (4N^3 + 3N^2)\,\text{RMUL}$, while puncturing requires $\theta_2 = \frac{2}{3}(8N^3 - 15N^2 + 4N - 12)\,\text{RADD} + (\frac{16}{3}N^3 - 7N^2 + \frac{8}{3}N - 20)\,\text{RMUL}$, assuming a $(1, (N-1)\times 1, N-1)$ structure. In comparison, [14] requires $(16N^4 - 4N^3)$ RMUL and $(16N^4 - 4N^3)$ RADD.

VI. Optimized Detection Algorithm

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

$$\pi_t(i) = \mathrm{mod}(i + (t-1)E - 1, N) + 1, i = 1, \cdots, N, \tag{20}$$

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 4 streams of a 4×4 MIMO system using Fig. 2(b), $T = 2$ 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. Subsets can overlap by altering π to place a low reliability stream in several detection sets. For simplicity, we assume $\mathcal{X}_n = \mathcal{O}$ for all n.

For detection, we enumerate all symbol vectors in $\mathcal{O}^E$ and then slice to find the closest symbols in the other S layers, populating the Euclidean distance of the resulting symbol vectors along the way. First partition $\tilde{\mathbf{y}}^{(t)}, \mathbf{L}^{(t)}$, and $\mathbf{x}$ as

$$\tilde{\mathbf{y}}^{(t)} = \begin{bmatrix} \tilde{\mathbf{y}}_1^{(t)} \\ \tilde{\mathbf{y}}_2^{(t)} \end{bmatrix}, \quad \mathbf{L}^{(t)} = \begin{bmatrix} \mathbf{A}^{(t)} & \mathbf{0} \\ tr\mathbf{B}^{(t)} & \mathbf{C}^{(t)} \end{bmatrix}, \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \tag{21}$$

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

$$\begin{aligned} \tilde{\mathbf{x}}_t^{\mathrm{WL}} &\triangleq \arg\min_{\mathbf{x}\in\mathcal{X}} \left\| \tilde{\mathbf{y}}^{(t)} - \mathbf{L}^{(t)}\mathbf{x} \right\|^2 \\ &= \arg\min_{\mathbf{x}_1\in\mathcal{O}^E} \left(\left\| \tilde{\mathbf{y}}_1^{(t)} - \mathbf{A}^{(t)}\mathbf{x}_1 \right\|^2 \right. \\ &\quad \left. + \left\| \tilde{\mathbf{y}}_2^{(t)} - \mathbf{B}^{(t)}\mathbf{x}_1 - \mathbf{C}^{(t)}\hat{\mathbf{x}}_2 \right\|^2 \right) \end{aligned} \tag{22}$$

$$\hat{\mathbf{x}}_2 = \left\lceil \left(\tilde{\mathbf{y}}_2^{(t)} - \mathbf{B}^{(t)}\mathbf{x}_1 \right) / \mathbf{C}^{(t)} \right\rfloor_{\mathcal{O}^s} \tag{23}$$

The first argument of the slicing operator in (23) is a vector of length S since $\mathbf{C}^{(t)}$ is a diagonal matrix. Slicing is applied to the individual vector elements over $\mathcal{O}$. The symbols in $\tilde{\mathbf{x}}_t^{\mathrm{WL}}$ are rearranged back to their normal order using (20), and the permuted vector is denoted as $\mathbf{x}_t^{\mathrm{WL}}$. The minimum distance d_t^{WL} itself is computed as the distance of $\mathbf{y}$ from $\mathbf{H}\mathbf{x}_t^{\mathrm{WL}}$ (rather than the distance of $\tilde{\mathbf{y}}^{(t)}$ from $\mathbf{L}^{(t)}\tilde{\mathbf{x}}_t^{\mathrm{WL}}$):

$$d_t^{\mathrm{WL}} = \left\| \mathbf{y} - \mathbf{H}\mathbf{x}_t^{\mathrm{WL}} \right\|^2 \neq \left\| \tilde{\mathbf{y}}^{(t)} - \mathbf{L}^{(t)}\tilde{\mathbf{x}}_t^{\mathrm{WL}} \right\|^2. \tag{24}$$

To compute the LLRs, we expand distances similar to (22) when taking $\arg\min$ to determine the symbol vector

$$\begin{aligned} \tilde{\mathbf{u}}_{n,k,t}^{\mathrm{WL}} &= \arg\min_{\mathbf{x}\in\mathcal{X}_{n,k}^{(0)}} \left\| \tilde{\mathbf{y}}^{(t)} - \mathbf{L}^{(t)}\mathbf{x} \right\|^2 \\ &= \arg\min_{\mathbf{x}_1\in\mathcal{O}_{n,k}^{(0)}} \left(\left\| \tilde{\mathbf{y}}_1^{(t)} - \mathbf{A}^{(t)}\mathbf{x}_1 \right\|^2 \right. \\ &\quad \left. + \left\| \tilde{\mathbf{y}}_2^{(t)} - \mathbf{B}^{(t)}\mathbf{x}_1 - \mathbf{C}^{(t)}\hat{\mathbf{x}}_2 \right\|^2 \right) \end{aligned} \tag{25}$$

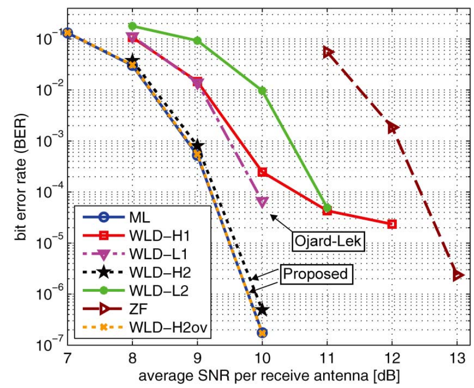


Fig. 3. BER versus SNR plots.

where $\mathcal{O}_{n,k}^{(0)} = \{\mathbf{x}_1 \in \mathcal{O}^E : b_{n,k} = 0\}$, and $\hat{\mathbf{x}}_2$ is defined as in (23) 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:

$$\tilde{\mathbf{v}}_{n,k,t}^{\mathrm{WL}} = \arg\min_{\mathbf{x}\in\mathcal{X}_{n,k}^{(1)}} \left\| \tilde{\mathbf{y}}^{(t)} - \mathbf{L}^{(t)}\mathbf{x} \right\|^2. \tag{26}$$

Then (20) is applied to permute the symbols in $\tilde{\mathbf{u}}_{n,k,t}^{\mathrm{WL}}$ and $\tilde{\mathbf{n}}_{n,k,t}^{\mathrm{WL}}$ and form the reordered symbol vectors $\mathbf{u}_{n,k,t}^{\mathrm{WL}}$ and $\mathbf{v}_{n,k,t}^{\mathrm{WL}}$. The LLR values are then computed as

$$\Lambda_{n,k,t}^{\mathrm{WL}} = \left\| \mathbf{y} - \mathbf{H}\mathbf{u}_{n,k,t}^{\mathrm{WL}} \right\|^2 - \left\| \mathbf{y} - \mathbf{H}\mathbf{v}_{n,k,t}^{\mathrm{WL}} \right\|^2 \tag{27}$$

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

One simple way to approximate the ML distance and LLRs is by selecting the minimum over all t in (24) and (27):

$$d^{\mathrm{ML}} \approx \min_{t=1,\cdots,T} \left(d_t^{\mathrm{WL}} \right), \qquad \Lambda_{n,k}^{\mathrm{ML}} \approx \min_{t=1,\cdots,T} \left(\Lambda_{n,k,t}^{\mathrm{WL}} \right). \tag{28}$$

Tighter LLRs are produced by tracking the *global minimum distances* across t rather than minimizing over the per stream LLRs. For $N = 2$, the algorithm is optimal and reduces to [11] since the $\mathbf{W}$'s are unitary. When $E = 1$ and distances in (24) are computed based on $\mathbf{L}$, the algorithm reduces to [14].

VII. Performance Simulation Results

The bit-error rate (BER) of the proposed detection algorithm is evaluated through simulations of a 4×4 MIMO system. The channel encoder is an LTE turbo encoder [16] with interleaver length 1024, and 16-QAM modulation is assumed. The channel entries are assumed to be i.i.d. complex Gaussian random variables. At the receiver, we assume perfect channel knowledge. The turbo decoder performs 4 decoding iterations. Fig. 3 compares the BER of the proposed WLD scheme with $E = 1$ and 2 structures, versus ML, ZF, and [14]. Both overlapping and non-overlapping subsets are considered. Two scenarios for distance computations in (24) are followed; one based on $\mathbf{H}$ and one on $\mathbf{L}$. The plots demonstrate that WLD with $E = 2$ using $\mathbf{H}$ distances with overlapping subsets performs virtually as ML, and is < 0.1 dB away from ML with no overlapping subsets. For single streams, $\mathbf{L}$ distances perform better than $\mathbf{H}$ distances.

REFERENCES

[1] A. Paulraj, R. Nabar, and D. Gore, *Introduction to space-time wireless communications*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[2] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1639–1642, Jul. 1999.

[3] O. Damen, A. Chkeif, and J.-C. Belfiore, "Lattice code decoder for space-time codes," *IEEE Commun. Lett.*, vol. 4, no. 5, pp. 161–163, May 2000.

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

[5] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Techn. J.*, vol. 1, no. 2, pp. 41–59, 1996, Autumn.

[6] Y. Jiang, J. Li, and W. W. Hager, "Joint transceiver design for MIMO communications using geometric mean decomposition," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3791–3803, Oct. 2005.

[7] Y. Jiang, J. Li, and W. W. Hager, "Uniform channel decomposition for MIMO communications," *IEEE Trans. Signal Process.*, vol. 53, no. 11, pp. 4283–4294, Nov. 2005.

[8] S. Ariyavisitakul, J. Zheng, E. Ojard, and J. Kim, "Subspace beamforming for near-capacity MIMO performance," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5729–5733, Nov. 2008.

[9] Y. Chen and S. Brink, "Near-capacity MIMO subspace detection," in *Proc. IEEE Int. Symp. Personal Indoor and Mobile Radio Commun. (PIMRC)*, Toronto, Canada, Sep. 2011, pp. 1733–1737.

[10] E. Viterbo and E. Biglieri, "A universal decoding algorithm for lattice codes," in *14ème Colloq. GRETSI*, Juan-Les-Pins, France, Sep. 1993, pp. 611–614.

[11] M. Siti and M. P. Fitz, "A novel soft-output layered orthogonal lattice detector for multiple antenna communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Istanbul, Turkey, Jun. 2006, vol. 4, pp. 1686–1691.

[12] M. Siti and M. P. Fitz, "On layer ordering techniques for near-optimal MIMO detectors," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Hong Kong, Mar. 2007, pp. 1199–1204.

[13] M. S. Yee, "Max-log-map sphere decoder," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Philadelphia, PA, USA, Mar. 2005, vol. 3, pp. 1013–1016.

[14] E. Ojard and S. Ariyavisitakul, "Method and system for approximate maximum likelihood (ML) detection in a multiple input multiple output (MIMO) receiver," U.S. Patent 12/207,721, Mar. 19, 2009 [Online]. Available: http://www.google.com/patents/US20090074114

[15] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd Ed. ed. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996.

[16] "Evolved universal terrestrial radio access (E-UTRA); Physical channels and modulation, 3GPP Std. TS 36.211," [Online]. Available: http://www.3gpp.org