Sequential Decoders for Large MIMO Systems

Konpal Shaukat Ali, Walid Abediseid, and Mohamed-Slim Alouini
Computer, Electrical and Mathematical Sciences and Engineering Division
King Abdullah University of Science and Technology, KAUST
Thuwal 23955-6900, Saudi Arabia
{konpal.ali, walid.abediseid, slim.alouini}@kaust.edu.sa

***Abstract*—Due to their ability to provide high data rates, multiple-input multiple-output (MIMO) systems have become increasingly popular. Decoding of these systems with acceptable error performance is computationally very demanding. In this paper, we employ the Sequential Decoder using the Fano Algorithm for large MIMO systems. A parameter called the bias is varied to attain different performance-complexity trade-offs. Low values of the bias result in excellent performance but at the expense of high complexity and vice versa for higher bias values. Numerical results are done that show moderate bias values result in a decent performance-complexity trade-off. We also attempt to bound the error by bounding the bias, using the minimum distance of a lattice. The variations in complexity with SNR have an interesting trend that shows room for considerable improvement. Our work is compared against linear decoders (LDs) aided with Element-based Lattice Reduction (ELR) and Complex Lenstra-Lenstra-Lovasz (CLLL) reduction.**

I. Introduction

The two goals of any communication system are achieving reliability and capacity, in other words, achieving high performance and high data rates, respectively. Multiple Input Multiple Output (MIMO) systems employ multiple antennas at the transmitter and receiver, which allow them to exploit the spatial dimension in order to achieve higher data rates and performance.

With the increasing demand of data rate in recent years, researchers have shown great interest in MIMO Systems. Unfortunately, a trade-off exists between achieving reliability and capacity. MIMO systems for instance, when used for improving data rates by sending different data streams on each transmit antenna - as opposed to sending the same data streams on multiple transmit antennas to achieve better performance - require very complex reliable detection methods, especially as the system size increases. The Maximum Likelihood (ML) decoder in particular, although optimal and therefore most reliable, suffers from exponential complexity in terms of the number of transmit antennas (M) and constellation size ($\mathcal{M}$). On the other hand, Linear Decoders (LDs) such as Zero-Forcing (ZF) and the Minimum Mean Square Error (MMSE) have only polynomial complexity and are thus widely adopted in a number of systems. In MIMO systems, however, these decoders result in very poor performance compared to the ML decoder due to their sensitivity to ill-conditioned channel matrices.

Analysis in [1] shows that for MIMO V-BLAST systems with M transmit antennas and N receive antennas, conventional LDs such as ZF and MMSE, can only collect a diversity of $N - M + 1$, though they enjoy very low computational complexity. The ML decoder, on the other hand, collects receive-diversity. However, Lattice Reduction (LR) techniques used to aid LDs, in [1], [2], [3], [4], [5], [6], [7] achieve receive-diversity at the expense of a small increase in complexity. It is important to note that a gap does exist between the performance curves of these LR-aided LDs and the ML detector due to the sub-optimality of the LDs, the imperfect orthogonalization of the channel matrix by lattice reduction and the imperfections introduced by the quantization operation. Other work such as [8] and [9] employ lattice reduction techniques as well, to achieve ML detection diversity. In our work, we employ the Sequential Decoder with the Fano Algorithm to not only achieve receive-diversity, but to decrease the gap that exists between the ML detectors performance and that of LR-aided LDs. By varying a parameter called the bias in the Sequential Decoder, we are able to attain a very good performance-complexity trade-off.

Our work will be compared against the Element-based Lattice Reduction (ELR) techniques used in [2] and the Complex Lenstra-Lenstra-Lovasz (CLLL) reduction technique employed in [1] to decode MIMO systems. The bias is varied according to the transmit-to-receive-antenna ratio in order to upper-bound the error probability according to [10]. Increasing the bias reduces complexity but results in higher error rates and vice versa for decreasing the bias. This happens because as the bias is increased, the decoder approaches the ZF decoder, while decreasing the bias results in it approaching the ML decoder. It is shown that the Sequential Decoder, like the CLLL and ELR, attains receive-diversity and, in fact, performs better than them.

The rest of the paper is organized as follows: in Section II, the MIMO system model is introduced. Section III explains past work done on decoding of MIMO systems using Lattice Reduction and discusses the ELR and CLLL techniques that our work will be compared against. Section IV describes the working of the Sequential Decoder, the framework of our system, and using the minimum distance of a lattice to upper-bound the error probability of the Sequential Decoder by varying a parameter called the bias accordingly. Section V contains the numerical results of simulations carried out using the decoding schemes for different systems. Section VI concludes the paper.

$Notation$: Throughout the paper, vectors will be denoted by lower bold faced letters and matrices by upper bold faced letters. The superscript ^{T} denotes the transpose, * denotes the conjugate and ^{H} denotes the Hermitian. Real and imaginary parts are denoted by $\Re[.]$ and $\Im[.]$ respectively. $\mathbf{I}_N$ denotes the $N \times N$ identity matrix and $\mathbf{1}_{N\times 1}$ the vector of dimension $N \times 1$ consisting of all ones. $\mathbb{Z}$ denotes the the integer set,

978-3-901882-63-0/2014 -

$\mathbb{C}$, the complex field, $\mathbb{R}$ for real numbers, and $\mathbb{Z}[j]$ for the Gaussian integer ring with elements of the form $\mathbb{Z} + j\mathbb{Z}$.

II. System Model

Our system is one that employs multiple antennas at both the transmitter and receiver. Let M denote the number of transmit antennas, N the number of receive antennas, and let $N \geq M$. The vectors $\mathbf{x}_c = [x_{c1}, x_{c2}, \ldots, x_{cM}]^T$, $\mathbf{y}_c = [y_{c1}, y_{c2}, \ldots, y_{cN}]^T$ and $\mathbf{w}_c = [w_{c1}, w_{c2}, \ldots, w_{cN}]^T$ denote the transmit, receive and noise vectors respectively. The $N \times M$ matrix $\mathbf{H}_{c_1}$ represents the channel matrix. The noise is assumed to be Additive White Gaussian Noise (AWGN) and the channel to undergo Rayleigh fading; each component of the noise vector and the channel matrix is independent identically distributed (i.i.d.) and has a complex Gaussian distribution with zero mean and unit variance. The channel is also assumed to be stationary throughout a transmission block and to vary independently from block to block. The Channel Side Information (CSI) is assumed to be available at the receiver but not at the transmitter. The following equation describes the channel model:

$$\mathbf{y}_c = \sqrt{\frac{D.SNR}{M}}\mathbf{H}_{c_1}\mathbf{x}_c + \mathbf{w}_c \tag{1}$$
$$= \mathbf{H}_c\mathbf{x}_c + \mathbf{w}_c, \tag{2}$$

where SNR is the signal to noise ratio, D is a normalizing factor equal to $12/(\mathcal{M}-1)$ and $\mathbf{H}_c = \sqrt{D.SNR/M}\mathbf{H}_{c_1}$. We have chosen the entries of $\mathbf{x}_c$ to be complex, independent and drawn from the QAM constellation $\mathcal{S}$ such that the real and imaginary parts of $\mathbf{x}_c$ are drawn from the set $\{\pm1, \pm3, \ldots, \pm\sqrt{\mathcal{M}}-1\}$.

The vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{w}$ are the vectors $\mathbf{x}_c$, $\mathbf{y}_c$ and $\mathbf{w}_c$ respectively, but with their real components stacked on top of their imaginary components i.e. $\mathbf{x} = [\Re[\mathbf{x}_c]^T \Im[\mathbf{x}_c]^T]^T$, $\mathbf{y} = [\Re[\mathbf{y}_c]^T \Im[\mathbf{y}_c]^T]^T$ and $\mathbf{w} = [\Re[\mathbf{w}_c]^T \Im[\mathbf{w}_c]^T]^T$. The counterpart of the matrix $\mathbf{H}_c$ is the matrix $\mathbf{H}$, where

$$\mathbf{H} = \begin{bmatrix} \Re[\mathbf{H}_c] & -\Im[\mathbf{H}_c] \\ \Im[\mathbf{H}_c] & \Re[\mathbf{H}_c] \end{bmatrix},$$

so that,

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}. \tag{3}$$

Note: $\mathbf{x}$ is of dimensions $2M \times 1$, $\mathbf{y}$ and $\mathbf{w}$ of $2N \times 1$ and $\mathbf{H}$ of $2N \times 2M$.

III. Lattice Reduction Techniques for Decoding MIMO Systems

Matrices that are more orthogonal have better properties such as ease of inversion, than matrices that are singular or ill-conditioned. The noiseless receive vector, $\mathbf{H}_c\mathbf{x}_c$, in MIMO detection can be interpreted as a point in a lattice with generator matrix $\mathbf{H}_c$. Since the same lattice can be generated by an infinite number of basis, different reduction techniques are available that result in a new generator matrix $\widetilde{\mathbf{H}}_c$ that has a condition number closer to 1, or in other words is more orthogonal.

$$\widetilde{\mathbf{H}}_c = \mathbf{H}_c\mathbf{T},$$

where $\mathbf{T}$ is a unimodular matrix and so, by definition, for the complex channel matrix $\mathbf{H}_c$, all elements of matrices $\mathbf{T}$ and $\mathbf{T}^{-1}$ contain Gaussian integers only and the determinant of $\mathbf{T}$ is ±1 or $\pm j$ (for real $\mathbf{H}_c$, all elements of matrices $\mathbf{T}$ and $\mathbf{T}^{-1}$ contain integers only and the determinant of $\mathbf{T}$ is 1).

We can now translate our original system in (2) to

$$\mathbf{y}_c = (\mathbf{H}_c\mathbf{T})(\mathbf{T}^{-1}\mathbf{x}_c) + \mathbf{w}_c \tag{4}$$
$$= \widetilde{\mathbf{H}}_c\widetilde{\mathbf{x}}_c + \mathbf{w}_c, \tag{5}$$

where $\widetilde{\mathbf{x}}_c = \mathbf{T}^{-1}\mathbf{x}_c$.

Linear detection schemes, such as ZF and MMSE, are applied to the modified system to obtain $\widehat{\widetilde{\mathbf{x}}}_c$. $\widehat{\widetilde{\mathbf{x}}}_c$ is then translated back by multiplying with $\mathbf{T}$ followed by quantization to obtain $\widehat{\mathbf{x}}_c$ belonging to the $\mathcal{M}$-QAM constellation, which is an estimate of $\mathbf{x}_c$.

$$\widehat{\mathbf{x}}_c = \mathcal{Q}(\mathbf{T}\widehat{\widetilde{\mathbf{x}}}_c). \tag{6}$$

The performance of these LR-aided LDs is significantly better than the LDs without any aid, and there is a small increase in the computational complexity of the LR-aided LDs. There are many techniques available to perform lattice reduction and in this paper we will compare our work to lattice reduction using the ELR algorithm and the CLLL algorithm constructed in [2] and [1] respectively.

A. Element-based Lattice Reduction (ELR)

It has been shown in [2] that the Pair-wise Error Probability (PEP) of incorrect detection of the i^{th} transmitted symbol, x_{c_i}, increases with the corresponding diagonal element $C_{i,i}$ of matrix $\mathbf{C}$, where $\mathbf{C} = ((\mathbf{H}_c^H)\mathbf{H}_c)^{-1}$. Similarly in the reduced basis the PEP of incorrect detection of the i^{th} transmitted symbol of $\widetilde{x}_{c_i}$ increases with the i^{th} diagonal element, $\widetilde{C}_{i,i}$, of matrix $\widetilde{\mathbf{C}}$ where

$$\widetilde{\mathbf{C}} = ((\widetilde{\mathbf{H}}_c^H)\widetilde{\mathbf{H}}_c)^{-1}$$
$$= \mathbf{T}^{-1}\mathbf{C}(\mathbf{T}^{-1})^H.$$

Thus to reduce the PEP, the diagonal elements of $\mathbf{C}$ must be minimized. In particular, increasing SNR results in the largest diagonal element of $\mathbf{C}$ dominating the PEP; minimizing this is therefore of utmost importance. Two optimization problems can be formulated from the above analysis:

- The Dual-Shortest Longest Vector (D-SLV) reduction, which minimizes the largest diagonal element of $\widetilde{\mathbf{C}}$ by finding an appropriate unimodular matrix
- The Dual-Shortest Longest Basis (D-SLB) reduction, which also by finding an appropriate unimodular matrix, minimizes each diagonal element of $\widetilde{\mathbf{C}}$ in descending order of value

Finding the optimum reduced-basis using the D-SLV and D-SLB is computationally very demanding. So instead ELR algorithms are proposed in [2] which iteratively calculate sub-optimal solutions for the D-SLV and D-SLB reductions.

B. Complex Lenstra-Lenstra-Lovasz (CLLL)

The LLL algorithm presented in [11] is one of the most popular techniques used for lattice reduction. Like the ELR algorithm, it does not guarantee finding the optimum reduced-basis as this may be computationally very complex, but instead guarantees to find a reduced basis within a factor of the optimum, in polynomial time. The LLL algorithm works with real matrices and hence the dimensionality of the channel matrices is doubled as computations are carried out on $\mathbf{H}$ instead of $\mathbf{H}_c$. The CLLL algorithm in [1], on the other hand, performs complex operations and can reduce complex matrices. As a result, the computational complexity of the CLLL algorithm is half of the LLL. Impressively, this reduction in complexity occurs without any performance loss.

IV. Sequential Decoder and related work

A. The Sequential Decoder using the Fano Algorithm

The procedure of lattice decoding basically involves the following: reducing the lattice basis to a more orthogonal basis, performing closest lattice point search (CLPS) in the reduced basis, and transforming the result back to the original basis.

In our work we employ the sub-optimal Sequential Decoder using the Fano Algorithm. Due to its sub-optimality, the decoder has a much lower computational complexity than the ML decoder but an error performance that, although acceptable, is worse than that of the ML. From [12] , the working of a Sequential decoder can be divided into two stages, namely the Preprocessing Stage followed by a Tree Search Stage. The purpose of the Preprocessing Stage is to tame the channel, make it sparser, and to put the problem in the form of a tree structure. Taming the channel involves a QR decomposition of the channel matrix so that the detection can be done recursively due to the upper triangular structure of the modified problem. MMSE-Decision Feedback Equalizer (MMSE-DFE) may also be applied to achieve better results at the decision point. To induce sparsity in the modified problem, a lattice reduction may be applied to obtain an upper triangular structure more sparse than the original one. Permutation of the columns of the upper triangular matrix also results in increased sparsity. To ensure the problem has a tree structure it must be in an upper triangular form. The Tree Search Stage involves finding the best path in the tree of possible codewords.

Since we are trying to avoid high complexity, we have not focused much on the Preprocessing Stage and simply apply a QR decomposition to the channel matrix to put our problem in the form of an upper triangular structure so that the problem has a tree structure and detection can be done recursively. For the ease of analysis we will denote the upper triangular system as follows:

$$\underbrace{\begin{bmatrix} z_{2M} \\ \vdots \\ z_1 \end{bmatrix}}_{\mathbf{z}} = \underbrace{\begin{bmatrix} r_{2M,2M} & \cdots & \cdots & r_{2M,1} \\ 0 & r_{2M-1,2M-1} & \cdots & r_{2M-1,1} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & r_{1,1} \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} s_{2M} \\ \vdots \\ s_1 \end{bmatrix}}_{\mathbf{s}} + \underbrace{\begin{bmatrix} n_{2M} \\ \vdots \\ n_1 \end{bmatrix}}_{\mathbf{n}},$$

where $\mathbf{z}$ represents the modified receive vector, $\mathbf{R}$ the modified channel, $\mathbf{s}$ the vector of symbols to be decoded, and $\mathbf{n}$ the modified noise vector.

This is followed by a Fano Search Stage which is a type of iterative search. The Fano Algorithm, unlike other search stage algorithms (e.g. the Stack Algorithm), requires almost negligible memory due its iterative nature. The price paid for the low memory requirements is the need to revisit nodes that the other algorithms do not. In particular, the Fano Algorithm is a Best-First Search algorithm which means that any level in the search stage the algorithm chooses the best possible child node and then checks the validity of the newly formed path. In general, the tree structure of a code is used to decode the received sequence by making tentative hypotheses on successive branches of the tree. These hypotheses may be changed when subsequent ones indicate an error in the previous hypotheses.

Nodes are represented by $\mathbf{s}^k$, where $1 \leq k \leq 2M$ and $\mathbf{s}^k = (s_1 s_2 \dots s_k)$, and the bias is denoted by b. Sequential decoding involves deciding a codeword that minimizes a certain function. The function is a bit metric called the path value and is denoted by $f(\mathbf{s}^k)$. The path value used in the Fano Algorithm of sequential decoding is

$$f(\mathbf{s}^k) = \sum_{j=1}^{k} w_j(\mathbf{s}^j) - bk,$$

where

$$w_i(\mathbf{s}^i) = \left| z_i - \sum_{j=1}^{i} r_{i,j} s_j \right|.$$

It has been proved that at any decoding stage, extending the path with the smallest Fano metric minimizes the probability that the extending path does not belong to the optimal path, thereby justifying the use of the path value. Making such a 'locally' optimal decision at every decoding stage, however, does not guarantee finding the 'global' optimal path (i.e. the ML detectors result), and hence the error performance of the Sequential Decoder using the Fano metric is inferior to ML decoding. The dynamic threshold, T, is another metric and is constrained to change in increments of a fixed number Δ, called the step size. The changes in the value of T are determined by the algorithm which tightens and loosens the bound T as required. Since the Fano Algorithm is a Best-First Search, only the child node with the best Fano Metric is considered in an iteration. If this metric is less than T, the child node is valid. For a valid child node the algorithm is terminated if the child node is a leaf node, otherwise tightening of the dynamic threshold T is done. If the child node isn't valid, the threshold will be increased if it is too small, and if it isn't the algorithm will move back a node and look for the next best child node.

The Sequential Decoder thus hypothesizes in such a way that the path value, $f(\mathbf{s}^k)$, is always less than the dynamic threshold T. If f is greater than T, and T is not too small, the decoder is on the wrong path and searching for a different path in the tree needs to be done.

The Fano Algorithm allows two types of movements from one node to another:
1. Forward $-$ the decoder goes one branch to the right in the received value tree from the previously hypothesized node
2. Backward $-$ moving one branch to the left in the received

value tree when an incorrect hypothesis has been made, so that the next best child node can be found.

A record of the previous, current, and successor nodes and path metrics is kept (i.e. $\mathbf{s}^{k-1}$, $f(\mathbf{s}^{k-1})$, $\mathbf{s}^k$, $f(\mathbf{s}^k)$, $\mathbf{s}^{k+1}$ and $f(\mathbf{s}^{k+1})$) and the threshold T at every node. Initially (i.e. at $k = 0$), $\mathbf{s}^k$ is the origin and $T = 0$.

B. Framework

Since the $2M \times 1$ vector $\mathbf{x}$ in (3) contains elements drawn from $\mathcal{M}$-QAM constellation i.e. $\{\pm 1, \pm 3, \ldots, \pm\sqrt{\mathcal{M}} - 1\}$, $\mathbf{x}$ can be rewritten as $\acute{\mathbf{x}}$, where $\mathbf{x} = 2\acute{\mathbf{x}} - \mathbf{1}_{2M\times 1}$, so that $\acute{\mathbf{x}}$ contains elements from $\mathbb{Z}^{2M}$. This translation is required as the Sequential Decoder works with integer vectors. The system in (3) can thus be expressed as

$$\begin{aligned} \mathbf{y} &= \mathbf{H}(2\acute{\mathbf{x}} - \mathbf{1}_{2M\times 1}) + \mathbf{w} && (7)\\ &= 2\mathbf{H}\acute{\mathbf{x}} - \mathbf{H}\mathbf{1}_{2M\times 1} + \mathbf{w}. && (8) \end{aligned}$$

The system can then be translated to

$$\begin{aligned} \acute{\mathbf{y}} &= \mathbf{y} + \mathbf{H}\mathbf{1}_{2M\times 1} && (9)\\ &= 2\mathbf{H}\acute{\mathbf{x}} + \mathbf{w} && (10)\\ &= \acute{\mathbf{H}}\acute{\mathbf{x}} + \mathbf{w}. && (11) \end{aligned}$$

where $\acute{\mathbf{H}} = 2\mathbf{H}$. The set $\Lambda = \{\acute{\mathbf{H}}\acute{\mathbf{x}} : \acute{\mathbf{x}} \in \mathbb{Z}^{2M}\}$ is a $2M$ dimensional lattice in $\mathbb{R}^{2N}$. Applying a QR-decomposition to $\acute{\mathbf{H}}$, we obtain $\mathbf{Q}_1$ and $\mathbf{R}_1$. This is used to calculate

$$\begin{aligned} \mathbf{y}_D &= \mathbf{Q}_1^H \acute{\mathbf{y}} && (12)\\ &= \mathbf{Q}_1^H \mathbf{Q}_1 \mathbf{R}_1 \acute{\mathbf{x}} + \mathbf{Q}_1^H \mathbf{w} && (13)\\ &= \mathbf{R}_1 \acute{\mathbf{x}} + \mathbf{w}_1, && (14) \end{aligned}$$

where $\mathbf{w}_1 = \mathbf{Q}_1^H \mathbf{w}$. As the columns of $\mathbf{Q}$ are unit vectors, the distribution of the elements of $\mathbf{w}_1$ is the same as that of $\mathbf{w}$. $\mathbf{y}_D$ is input to the Fano Algorithm in [12] along with $\mathbf{R}_1$, the step size which is set to be equal to 1 for our work, and the bias. The bias is the parameter which is varied to obtain different performance-complexity trade-offs. The output of the Fano Algorithm is the $2M \times 1$ vector $\widehat{\acute{\mathbf{x}}}$, consisting of integers. It is then translated back to obtain the $2M \times 1$ vector $\widehat{\mathbf{x}}_1$ by the following

$$\widehat{\mathbf{x}}_1 = 2\widehat{\acute{\mathbf{x}}} - \mathbf{1}_{2M\times 1}.$$

The vector $\widehat{\mathbf{x}}_1$ consists of the real and imaginary parts stacked on top of one another, these are then used to obtain the complex vector $\widehat{\mathbf{x}}_{c1}$ of dimensions $M \times 1$. Since the Fano Algorithm outputs a vector containing integers from the infinite ring whereas the original transmitted vector contains elements belonging to the $\mathcal{M}$-QAM constellation, a quantization step is required to ensure the decoded symbols belong to the $\mathcal{M}$-QAM constellation. Hence $\widehat{\mathbf{x}}_c$, which is an estimate of $\mathbf{x}_c$ in (2), is obtained by the quantization of $\widehat{\mathbf{x}}_{c1}$, i.e. $\widehat{\mathbf{x}}_c = \mathcal{Q}(\widehat{\mathbf{x}}_{c1})$.

C. Minimum Distance of a Lattice

In this section we will discuss how the minimum euclidean distance of a lattice generated by the $n\times m$ real channel matrix, $\mathbf{H}$, is used to bound error probability by bounding the bias. Assuming the all-zero lattice point of dimensions $m \times 1$ is transmitted, the error probability of the Sequential Decoder can be upper bounded as a function of the bias as shown in Eqn. (9) of [10]

$$P_e(b) \leq Pr\left(\bigcup_{\substack{\mathbf{x}\in\mathbb{Z}^m\\ \mathbf{x}\neq\mathbf{0}}} \left\{ 2\mathbf{x}^T\mathbf{w} > ||\mathbf{x}||^2\Big(1 - \frac{bm}{d_{min}^2(\mathbf{H})}\Big)\right\} \right).$$

From this we can see that we require the term $bm/d_{min}^2(\mathbf{H})$ to be less than 1. Hence, the bias, b, must be lower than $d_{min}^2(\mathbf{H})/m$, where $d_{min}^2(\mathbf{H})$ is defined as

$$d_{min}^2(\mathbf{H}) = \min_{\substack{\mathbf{x}\in\mathbb{Z}^m\\ \mathbf{x}\neq\mathbf{0}}} ||\mathbf{H}\mathbf{x}||^2.$$

Thus we are interested in finding the minimum distance of the noiseless received vector, obtained after multiplication of an integer transmit vector $\mathbf{x}$ with the channel matrix $\mathbf{H}$. As $\mathbf{x}$ can be any point on the m-dimensional integer space and the integer space has infinite points, using an exhaustive search such as ML decoding to find the minimum distance is not possible. Hence for the calculation of $d_{min}(\mathbf{H})$ we opt for a technique that has lower computational complexity, namely the Sphere Decoder presented in [13]. Note, as $\mathbf{H}$ is random, an average minimum distance can be found by averaging over a large number of realizations of the channel matrix. This average minimum distance is used for calculating the bias in our work.

For a particular ratio of transmit to receive antennas, $\ddot{y} = m/n$, increasing the number of transmit antennas and hence the number of receive antennas as well, an asymptotic bound on the minimum and maximum eigenvalues of channel matrices was found in [14]. The ratio of the minimum eigenvalue to the number of transmit antennas, λ_{min}/m, of an $n \times m$ channel matrix approaches $2(1 - \sqrt{\ddot{y}})^2$ as the number of antennas is increased for a fixed transmit to receive antenna ratio $\ddot{y}$. Additionally $d_{min}^2(\mathbf{H})$ is lower bounded by λ_{min} as:

$$\begin{aligned} d_{min}^2(\mathbf{H}) &= \min_{\substack{\mathbf{x}\in\mathbb{Z}^m\\ \mathbf{x}\neq\mathbf{0}}} ||\mathbf{H}\mathbf{x}||^2\\ &\geq \min_{\substack{\mathbf{x}\in\mathbb{Z}^m\\ \mathbf{x}\neq\mathbf{0}}} \mathbf{x}\mathbf{H}^T\mathbf{H}\mathbf{x}\\ &\geq \lambda_{min} \min_{\substack{\mathbf{x}\in\mathbb{Z}^m\\ \mathbf{x}\neq\mathbf{0}}} ||\mathbf{x}||^2\\ &= \lambda_{min}. \end{aligned}$$

Note:

$$\min_{\substack{\mathbf{x}\in\mathbb{Z}^m\\ \mathbf{x}\neq\mathbf{0}}} ||\mathbf{x}||^2 = 1,$$

as the smallest norm corresponds to the vector with all entries equal to 0, except one entry which is equal to 1.

Through simulations using the Sphere Decoder we were able to calculate and plot $d_{min}^2(\mathbf{H})/m$ for different antenna ratios $\ddot{y}$ with increasing antennas. We require this as we want to investigate the effect of using bias values in the range of the lowest $d_{min}^2(\mathbf{H})/m$ for a particular $\ddot{y}$.

V. Numerical Results

In this section we compare the performance and complexity of Sequential Decoders with and without an MMSE extended system, using different bias values, against LR-aided ZF and

MMSE LDs. The LR techniques employed for comparison are the dual ELR SLB and the CLLL. The bias values used for the Sequential Decoder are 0.4, 0.5, 1 and 5 so that the performance-complexity trade-off can be observed. All the systems being analyzed contain equal numbers of transmit and receive antennas. The bias values of 0.4 and 0.5 were chosen as the values in the range lower than $d^2_{min}(\mathbf{H})/m$ for $\ddot{y} = 1$.

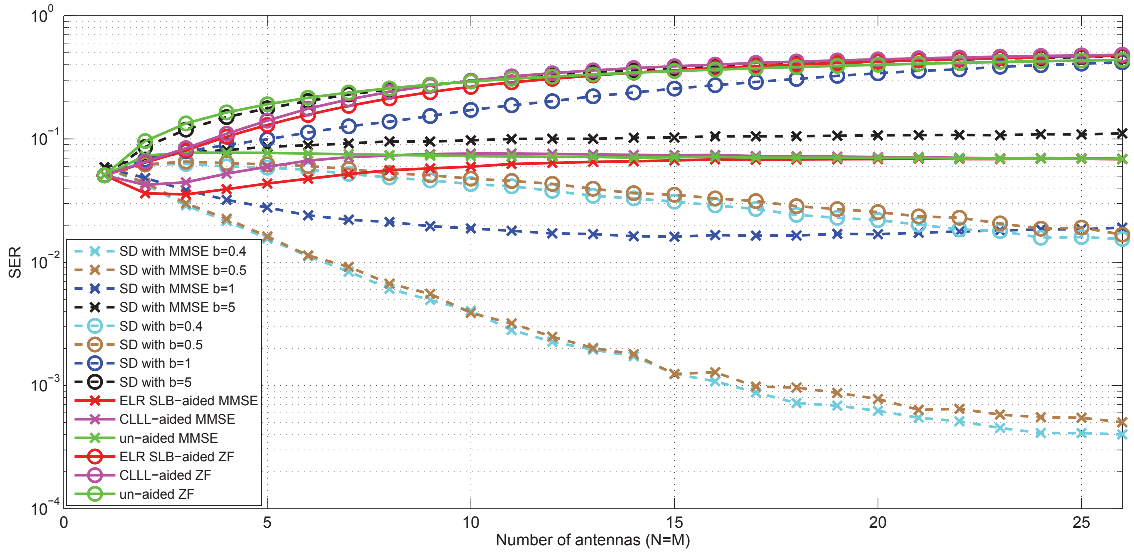
Fig. 1(a) is a plot of the symbol error rate as the number of antennas is increased. A 4-QAM constellation is used for transmission and an SNR of 3dB. As expected, the un-aided ZF decoder has the worst performance. The CLLL-aided ZF shows some improvement and the ELR SLB-aided ZF has the best performance among the LR aided ZF decoders. The Sequential Decoder with a bias of 5 performs not much better than the unaided ZF, but at a bias of 1 there is a marked improvement in performance, and bias values of 0.4 and 0.5 show particularly good performance. In the cases of bias values equal to 0.4 and 0.5, it is also very interesting to note that unlike the other decoders, performance is decreasing with increasing number of antennas. The LR aided MMSE and unaided MMSE showed marked improvement but similar performance trends to their ZF counterparts. The extended MMSE system was also applied to the Sequential Decoders and performance improvement was again noted from the corresponding un-extended case; the performance with bias values equal to 0.4 and 0.5 is exceptional.

Fig. 1(b) shows the corresponding complexities of the Sequential Decoders and the complexity of the lattice reduction for each technique applied to both the channel matrix (ZF case) and the extended channel matrix (MMSE case). As expected, the complexity of the MMSE systems is lower than their ZF counterparts and the complexities grow with increasing antennas. Among the LR schemes, ELR SLB reduction has the highest complexity, followed by CLLL. The rate at which complexity for the Sequential Decoders increases is larger than that of the lattice reductions particularly for the smaller bias values. The complexity of the Sequential Decoder with the extended MMSE system shows considerable improvement and is even lower than the SLB-aided MMSE for smaller systems, but when the number of antennas exceeds around 12 to 20, the complexity of the Sequential Decoder with lower bias values becomes larger. The price paid for the performance gain of the Sequential Decoder is its high complexity for large systems. For small systems, such as of five antennas, it is interesting to note that there is a performance gain without higher complexity.

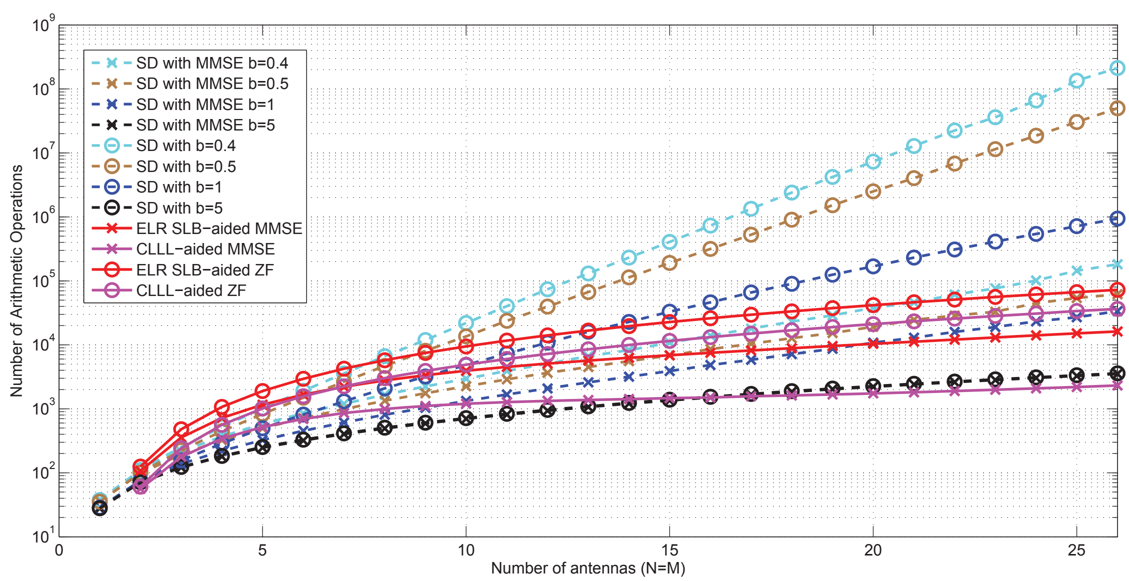
Fig. 2(a) is the plot of the symbol error versus SNR of a 32×32 system employing a 16-QAM constellation. As can be seen from the figure, the performance of the Sequential Decoder is superior and its error curves fall a lot more sharply and at considerably lower SNR values. The error curves corresponding to lower bias values fall at lower SNR and the error curves fall almost parallel to one another. Additionally, the Sequential Decoder when applied to an extended MMSE system results in the error curves falling at even lower SNR values, and the gap can be seen between them and their corresponding un-extended system. The ELR SLB-aided ZF and MMSE have the next best performance after the Sequential Decoders, though the ELR SLB-aided ZF has better performance at moderate SNR values and at higher SNR the ELR SLB-aided MMSE outperforms it. The CLLL-aided MMSE and ZF perform next best. At low to moderate SNR, the CLLL-aided MMSE outperforms the CLLL-aided ZF, but at an SNR of around 20dB, the CLLL-aided ZF's error falls and it outperforms its MMSE counterpart. The worst performance as expected is by the unaided ZF, followed by the unaided MMSE.

Fig. 2(b) is a plot of the corresponding complexity of the Sequential Decoders with and without the extended MMSE system, and the Lattice Reduction techniques applied to the channel matrix and the extended MMSE channel matrix. The complexity for Lattice Reduction of the un-extended ZF systems is constant and independent of SNR; the ELR SLB having the highest complexity, followed by the CLLL. The corresponding MMSE systems have lower complexity at low SNR values, but as SNR increases the complexity of the MMSE system's Lattice Reduction reaches that of the ZF system's. This can be explained by the fact that in the case of the MMSE, the channel matrix is replaced by the extended channel matrix, which is better conditioned. Since both the LR algorithms perform reduction until the matrix is good enough in terms of orthogonality of the columns, the extended channel matrix in a lot of instances does not require reduction as it is already good enough. When the unit variance channel matrix is multiplied by a larger factor, the need to reduce the channel matrix and make it closer to what is defined as 'minimal orthogonality' is increased. Increasing SNR therefore results in more channel matrices undergoing reduction. This increases until all the channel matrices need to be reduced and hence we have constant complexity for the higher SNR values. This can also be seen by varying a parameter of the CLLL reduction algorithm, δ, which is a measure of how much the channel matrix should be reduced to make it close to orthogonal. Increasing δ results in more reduced matrices, higher computational complexity and better performance. δ can take any value between 0.5 and 1, and was set to 0.5 for all of our simulations.

The Sequential Decoders in Fig. 2(b) have a particular complexity trend: a region of low complexity (comparable to the LR techniques) at low SNR, a region of high complexity in a particular SNR range, low complexity (comparable to the LR techniques) at higher SNR. The range of high complexity shifts towards higher SNR values as the bias is increased. It should be noted that for each bias value the range of SNR that has high complexity corresponds to the SNR range in the error performance curves where the error probability first starts to drop steeply. After this high-complexity range, at higher SNR values the complexity of the Sequential Decoders falls to values comparable and even lower than the LR techniques. These trends can be explained as follows; initially at low SNR as the noise is more significant relative to the signal part of the received message, the decoder finds incorrect branches on the tree and makes mostly erroneous decisions and without much effort, corresponding to the high error probability and low complexity. As the SNR is increased, the decoder is able to differentiate between the noise and signal components so the complexity increases and error decreases as more work is done to decode correctly by finding the right branch. At higher SNR values, it is easier for the decoder to differentiate between the error and signal components and correct detection is done without too much work by the decoder, hence low

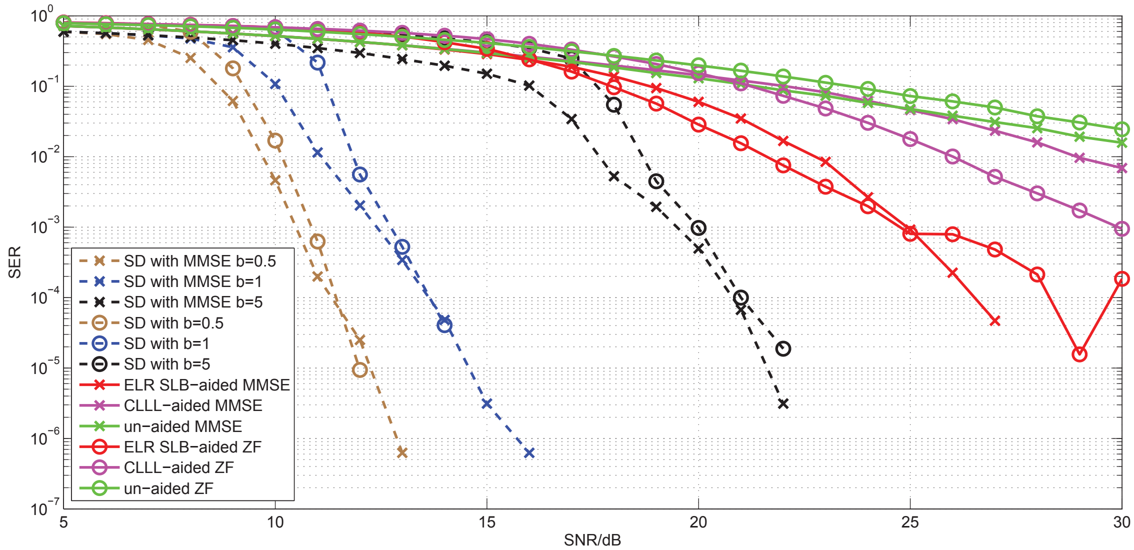


(a) Performance of different detectors for increasing number of antennas.

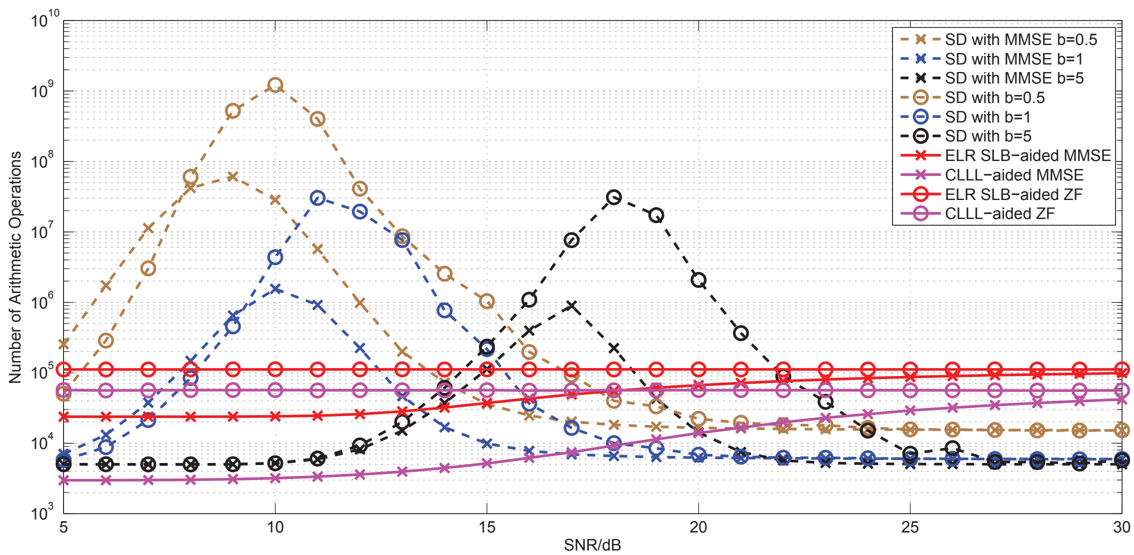


(b) Complexity of different detectors for increasing number of antennas.

Fig. 1: Performance and complexity of different detectors for MIMO Systems employing 4-QAM, SNR=3dB and equal number of transmit and receive antennas.



(a) Performance vs. SNR for different detectors.



(b) Complexity vs. SNR for different detectors.

Fig. 2: Performance and complexity of different detectors for a 32×32 MIMO System employing 16-QAM.

error at low complexity. It should also be noted that the Sequential Decoders applied to the extended MMSE systems, although still have higher complexity than the Lattice Reduction techniques, show a significant complexity improvement from their un-extended counterparts, have a narrower range of high-complexity and this range occurs a little before the that of the un-extended Sequential Decoders which corresponds to the fact that their error curves drop before those of the un-extended.

VI. Conclusion

In this paper, we analyzed the performance of Sequential Decoders for large MIMO systems and compare them to LR-aided LDs. The results show that the Sequential Decoder outperforms LR-aided LDs in terms of performance, though complexity may be high. It can also be extrapolated that there is room for improvement of the overall system performance, as the bias and SNR for a particular number of transmit and receive antennas can be chosen so that significant performance improvements are achieved in the same or even lower complexity range.

REFERENCES

[1] X. Ma and W. Zhang, “Performance analysis for MIMO systems with lattice-reduction aided linear equalization,” *IEEE Trans. Commun.*, vol. 56, no. 2, pp. 309-318, Feb. 2008.

[2] Q. Zhou and X. Ma, “Element-based lattice reduction algorithms for large MIMO detection,” *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 274-286, Feb. 2013.

[3] H. Yao and G. W. Wornell, “Lattice-reduction-aided detectors for MIMO communication systems,” in *Proc. IEEE Global Tele. Conf.*, Nov. 2002, pp. 424-428.

[4] C. Windpassinger and R. Fischer, “Low-complexity near-maximumlikelihood detection and precoding for MIMO systems using lattice reduction,” in *Proc. IEEE Info. Theory Workshop.* Paris, France, Mar. 2003, pp. 345-348.

[5] Y. H. Gan, C. Ling, and W. H. Mow, “Complex lattice reduction algorithm for low-complexity full diversity MIMO detection,” *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2701-2710, Jul. 2009.

[6] M. Taherzadeh, A. Mobasher, and A. Khandani, “LLL reduction achieves the receive diversity in MIMO decoding,” *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4801-4805, Dec. 2007.

[7] C. Chen and W. Sheen, “A new lattice reduction algorithm for LRaided MIMO linear detection,” *I Trans. Wireless Commun.*, vol. 10, no. 99, pp. 1-6, Aug. 2011.

[8] K. A. Singhal, T. Datta, and A. Chockalingam, “Lattice reduction aided detection in large-MIMO systems,” *2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 594-598, Jun. 2013.

[9] Q. Zhou and X. Ma, “Joint transceiver designs using lattice reduction algorithms,” *2013 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*, pp. 584-588, Jul. 2013.

[10] W. Abediseid and M. S. Alouini, “On lattice sequential decoding for the unconstrained AWGN channel,” *IEEE Trans. on Comm.*, vol. 61, no. 6, pp. 2446-2456, Jun. 2013

[11] A. K. Lenstra, H. W. Lenstra, and L. Lovasz, “Factoring polynomials with rational coefficients,” *Math. Annalen*, vol. 261, no. 4, pp. 515-534, 1982.

[12] A.D. Murugan, H. El Gamal, M.O. Damen and Caire, “A unified framework for tree search decoding: Rediscovering the sequential decoder,” *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 933-953, Mar. 2006.

[13] M.O. Damen, H. El Gamal and G. Caire, “On maximum-likelihood detection and the search for the closest lattice point,” *IEEE Trans. Inf. Theory* , vol. 49. no. 10, pp. 2389-2402, Oct. 2003.

[14] A. Edelman, “Eigenvalues and condition numbers of random matrices”, M.I.T. Doctoral Dissertation, Mathematics Department, 1989