Truncation for Low Complexity MIMO Signal Detection

Wen Jiang and Xingxing Yu
School of Mathematics
Georgia Institute of Technology, Atlanta, Georgia, 30332
Email: wjiang@math.gatech.edu, yu@math.gatech.edu

Ye (Geoffrey) Li
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, Georgia, 30332
Email: liye@ece.gatech.edu

***Abstract*–Joint maximum-likelihood (JML) detector may be used in memoryless multiple input multiple output (MIMO) systems to obtain optimal detection performance. However, the JML detector needs an exhaustive search and causes prohibitively large decoding complexity. To reduce the complexity of MIMO signal detection, minimum mean-square-error (MMSE) linear detector (LD), decision-feedback detector (DFD) and sphere detector (SD) may be used. In this paper, we propose a truncation based detector for low complexity MIMO signal detection, and give theoretical insight into the design and performance. We study bi-truncation in detail and present two bi-truncation approaches. These approaches have low-complexity, and computer simulation results show that they outperform MMSE-LD and MMSE-DFD.**

***Keywords*–MIMO signal detection, channel truncation, bi-truncation, Viterbi Algorithm**.

I. INTRODUCTION

In MIMO systems, *joint maximum-likelihood* (JML) detector minimizes the joint error probability. However, its complexity increases exponentially as the number of input bits increases, which is often impractical. To reduce decoding complexity, some simplified detection approaches have been employed, including *linear detector* (LD) [1], *decision-feedback detector* (DFD) [2], sorted DFD [3], and *sphere detector* (SD) [4], [5]. In this paper, we present an original truncation based detector for simplified MIMO signal detection. Truncation detector uses a linear (matrix) transformation to truncate the channel into an L-diagonal matrix. Then slightly modified Viterbi Algorithm [6], [7], [8] is used to detect the signals with low complexity.

The rest of the paper is organized as follows. In Section II, we briefly describe a MIMO system with a truncation based signal detector. Then in Section III, we derive truncation criteria in terms of maximizing *signal-to-noise ratio* (SNR) and minimizing error probability bound. As an example, we study bi-truncation in detail. The specific bi-truncation criteria are investigated in Section IV. Based on such truncation criteria, we develop two bi-truncation approaches in Section V and investigate the optimal grouping in Section VI. Finally, we present computer simulation results and compare with some some known detectors to demonstrate effectiveness of the proposed methods in Section VII and conclude the paper in Section VIII.

II. SYSTEM MODEL

A memoryless m-input and n-output channel is described in Figure 1. The complex received signal vector, $\mathbf{r} = (r_1, r_2, \cdots, r_n)^T$, can be expressed as

$$\mathbf{r} = \sqrt{E_s}\mathbf{H}\mathbf{a} + \mathbf{z}, \tag{1}$$

This work is partially supported by the NSA under Grant MDA904-03-1-0052 and NSF under Grant DMS-0245530.

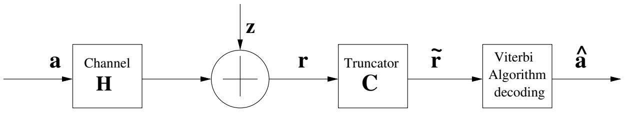

Fig. 1. A MIMO system with a truncation detector.

where $\mathbf{H}$ is an $n \times m$ channel matrix, $\mathbf{a} = (a_1, a_2, \cdots, a_m)^T$ is the channel input vector, and $\mathbf{z} = (z_1, z_2, \cdots, z_n)^T$ is the noise vector.

The m channel input symbols are independent and uniformly chosen from the same discrete and complex alphabet set. Therefore, input symbols, a_i's, are *independent and identically distributed* (i.i.d), and with zero-mean and unit-variance. We assume that the m columns of $\mathbf{H}$ are linearly independent, which requires that there are at least as many outputs as inputs, that is, $m \le n$. As usual, we also assume that the noise is white (circular complex) Gaussian so that the real and imaginary components of $\mathbf{z}$ are i.i.d zero-mean Gaussian with variance $N_0/2$. In addition, we make the assumption that the transmitter does not know $\mathbf{H}$, but the receiver has perfect channel knowledge.

We desire to find a linear truncator $\mathbf{C}$ such that $\mathbf{B} = \mathbf{CH}$ is an L-diagonal matrix as in (2). If $L = 2$, we call this truncation *bi-truncation*. Such a matrix $\mathbf{C}$ always exists if the columns of $\mathbf{H}$ are linearly independent. After the truncator, $\mathbf{C}$, the output becomes

$$\tilde{\mathbf{r}} = \sqrt{E_s}\mathbf{Ba} + \tilde{\mathbf{z}}, \tag{3}$$

where the noise vector $\tilde{\mathbf{z}} = \mathbf{Cz} = (\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_m)^T$. Note that $\tilde{\mathbf{z}}$ is usually not white any more, but it is still Gaussian and with zero-mean and autocorrelation matrix $N_0\mathbf{CC}^*$. After the channel is truncated in this form, a slightly modified Viterbi Algorithm is then used to detect the transmitted symbols. For bi-truncation, the number of states in the trellis diagram is the same as the number of constellations. For 4-QAM constellation, there are only 4 states; therefore, the complexity is much less than that of the JML detector.

III. TRUNCATION CRITERIA

In this section, we shall investigate truncation criteria to minimize error probability bounds or maximize the average SNR.

A. Maximum of Average SNR

Under the assumptions that $\mathbf{z}$ and $\mathbf{a}$ are i.i.d., with zero-mean, and variance N_0 and 1, respectively, the average SNR is

$$\begin{aligned} \text{SNR} &= \frac{E(|\sqrt{E_s}\mathbf{Ba}|^2)}{E(|\tilde{\mathbf{z}}|^2)} \\ &= \frac{E_s}{N_0}\frac{\text{tr}(\mathbf{BB}^*)}{\text{tr}(\mathbf{CC}^*)}. \end{aligned} \tag{4}$$

Our objective here is to find $\mathbf{B}$ and $\mathbf{C}$ to maximize SNR. From singular value decomposition theory, there exist an $n \times n$ unitary matrix, $\mathbf{U}$, and an $m \times m$ unitary matrix, $\mathbf{V}$, such that

$$\mathbf{H} = \mathbf{U}^* \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_m \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{V},$$

where $\lambda_1 \ge 0$, $\lambda_2 \ge 0$, $\cdots$, $\lambda_m \ge 0$ are the singular values of $\mathbf{H}$. Since we assume the columns of $\mathbf{H}$ are linearly independent, $\lambda_i > 0 \ \text{ for } \ 1 \le i \le m.$

Let $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m)$ and $\mathbf{CU}^* = (\mathbf{C}_1, \mathbf{C}_2)$, where $\mathbf{C}_1$ is an $m \times m$ matrix, and $\mathbf{C}_2$ is an $m \times (n - m)$ matrix. Since $\mathbf{CH} = \mathbf{B}$, we immediately get $\mathbf{C}_1 = \mathbf{BV}^*\mathbf{D}^{-1}$. Hence,

$$\mathbf{C} = (\mathbf{C}_1 \ , \ \mathbf{C}_2)\mathbf{U} = (\mathbf{BV}^*\mathbf{D}^{-1} \ , \ \mathbf{C}_2)\mathbf{U} \tag{5}$$

and

$$\begin{aligned} \mathbf{CC}^* &= \mathbf{BV}^*\mathbf{D}^{-1}(\mathbf{D}^{-1})^*\mathbf{VB}^* + \mathbf{C}_2\mathbf{C}_2^* \\ &= \mathbf{BXB}^* + \mathbf{C}_2\mathbf{C}_2^*, \end{aligned} \tag{6}$$

where $\mathbf{X} = (\mathbf{V}^*\mathbf{D}^{-1})(\mathbf{V}^*\mathbf{D}^{-1})^*$ is an $m \times m$ positive definite Hermitian matrix since both $\mathbf{D}$ and $\mathbf{V}$ are nonsingular. From (6), we have

$$\text{tr}(\mathbf{CC}^*) \ge \text{tr}(\mathbf{BXB}^*), \tag{7}$$

and this equality holds if and only if $\mathbf{C}_2 = \mathbf{0}$. From (4) and (7), we obtain the following Theorem.

Theorem 1 *The total average SNR of system (3) is bounded by*

$$\text{SNR} \le \frac{E_s}{N_0}\frac{\text{tr}(\mathbf{BB}^*)}{\text{tr}(\mathbf{BXB}^*)}, \tag{8}$$

$$\mathbf{B} = \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,L} & 0 & 0 & \cdots & 0 \\ 0 & d_{2,1} & d_{2,2} & \cdots & d_{2,L} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & d_{i-1,1} & d_{i-1,2} & \cdots & d_{i-1,L} & 0 \\ 0 & 0 & 0 & 0 & d_{i,1} & d_{i,2} & \cdots & d_{i,L} \\ d_{i+1,L} & 0 & \ddots & 0 & 0 & d_{i+1,1} & \cdots & d_{i+1,L-1} \\ d_{i+2,L-1} & d_{i+2,L} & 0 & \ddots & 0 & 0 & d_{i+2,1} & \cdots \\ \vdots & \ddots & 0 & \ddots & \ddots & 0 & \ddots & \vdots \\ d_{m,2} & \cdots & d_{m,L} & 0 & 0 & \cdots & 0 & d_{m,1} \end{pmatrix} \tag{2}$$

and the maximum is achieved if and only if

$$\mathbf{C} = (\mathbf{B}\mathbf{V}^*\mathbf{D}^{-1}, \mathbf{0}), \tag{9}$$

that is equivalent to $\mathbf{C}_2 = \mathbf{0}$.

From Theorem 1, we conclude that a good approach to maximizing SNR is therefore to minimize tr$(\mathbf{B}\mathbf{X}\mathbf{B}^*)$ while fixing tr$(\mathbf{B}\mathbf{B}^*)$ and setting $\mathbf{C}_2 = \mathbf{0}$.

B. Error Probability Analysis

If a signal $\mathbf{r}$ is detected by JML detector from (1), then the pairwise error probability of transmitting $\mathbf{a} = (a_1, a_2, \cdots, a_m)$ and decoding in favor of $\mathbf{e} = (e_1, e_2, \cdots, e_m)$ is well approximated [9], [10] by

$$P_{\mathbf{a}\to\mathbf{e}} \le exp(-|\mathbf{Ha} - \mathbf{He}|^2 \frac{E_s}{4N_0}). \tag{10}$$

In our scheme, a truncator $\mathbf{C}$ is used at the receiver, and equation (3) is used to detect the signal. Based on the approach presented in [9], we now derive an upper bound on the error probability by Markov's inequality [11]. Assuming Viterbi Algorithm is used for decoding, then the pairwise error probability is

$$P_{\mathbf{a}\to\mathbf{e}} = Pr(|\tilde{\mathbf{r}} - \sqrt{E_s}\mathbf{Ba}|^2 \ge |\tilde{\mathbf{r}} - \sqrt{E_s}\mathbf{Be}|^2). \tag{11}$$

For simplicity, we set $\boldsymbol{\eta} := \mathbf{Ba} - \mathbf{Be}$ and write $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_m)^T$. From (3), we find

$$|\tilde{\mathbf{r}} - \sqrt{E_s}\mathbf{Ba}|^2 = |\tilde{\mathbf{z}}|^2,$$

and

$$|\tilde{\mathbf{r}} - \sqrt{E_s}\mathbf{Be}|^2 = E_s|\boldsymbol{\eta}|^2 + 2\sqrt{E_s}Re(\boldsymbol{\eta}^*\tilde{\mathbf{z}}) + |\tilde{\mathbf{z}}|^2,$$

where Re$(\cdot)$ denotes the real part of a complex number. So (11) can be rewritten as

$$P_{\mathbf{a}\to\mathbf{e}} = Pr(-2\sqrt{E_s}Re(\boldsymbol{\eta}^*\tilde{\mathbf{z}}) \ge E_s|\boldsymbol{\eta}|^2). \tag{12}$$

Let $\lambda > 0$ be arbitrary. Note that $Y \ge \alpha$ if and only if $e^{\lambda Y} \ge e^{\lambda\alpha}$. Applying Markov's inequality we have

$$\begin{aligned} P_{\mathbf{a}\to\mathbf{e}} = & \; Pr(exp\{-2\lambda\sqrt{E_s}Re(\boldsymbol{\eta}^*\tilde{\mathbf{z}})\} \ge exp\{\lambda E_s|\boldsymbol{\eta}|^2\}) \\ \le & \; \frac{E(exp\{(-2\lambda\sqrt{E_s})Re(\boldsymbol{\eta}^*\tilde{\mathbf{z}})\})}{exp\{\lambda E_s|\boldsymbol{\eta}|^2\}}. \end{aligned} \tag{13}$$

Recall that $\tilde{\mathbf{z}} = \mathbf{Cz} = (\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_m)^T$. Hence

$$\begin{aligned} Re(\boldsymbol{\eta}^*\tilde{\mathbf{z}}) = & \; \textstyle\sum_{i=1}^m Re(\eta_i^*\tilde{z}_i) \\ = & \; \textstyle\sum_{k=1}^m Re(z_k)(\sum_{i=1}^m Re(\eta_i\mathbf{C}_{i,k}^*)) \\ & + \textstyle\sum_{k=1}^m Im(z_k)(\sum_{i=1}^m Im(\eta_i\mathbf{C}_{i,k}^*)), \end{aligned}$$

where Im$(\cdot)$ denotes the imaginary part of a complex number. Under the assumption that the real and imaginary components of $\mathbf{z}$ are i.i.d zero-mean Gaussian with variance $N_0/2$, the variance of $Re(\boldsymbol{\eta}^*\tilde{\mathbf{z}})$ is $\sigma^2_{Re(\boldsymbol{\eta}^*\tilde{\mathbf{z}})} = \frac{N_0}{2}\sigma^2$, where

$$\sigma^2 = \sum_{k=1}^m((\sum_{i=1}^m Re(\eta_i\mathbf{C}_{i,k}^*))^2 + (\sum_{i=1}^m Im(\eta_i\mathbf{C}_{i,k}^*))^2). \tag{14}$$

Note that for any real number t, the characteristic function of a Gaussian random variable Y with zero-mean and variance σ_Y^2 is $E(e^{jtY}) = exp\{-\frac{t^2\sigma_Y^2}{2}\}$. Hence from (13) we have

$$\begin{aligned} P_{\mathbf{a}\to\mathbf{e}} \le & \; \frac{exp\{\frac{1}{2}(-2\lambda\sqrt{E_s})^2\sigma^2_{Re(\boldsymbol{\eta}^*\tilde{\mathbf{z}})}\}}{exp\{\lambda E_s|\boldsymbol{\eta}|^2\}} \\ = & \; exp\{E_s(N_0\sigma^2\lambda^2 - |\boldsymbol{\eta}|^2\lambda)\}. \end{aligned} \tag{15}$$

Optimizing (15), we have $\lambda = \frac{|\boldsymbol{\eta}|^2}{2N_0\sigma^2}$, which yields the following upper bound:

$$P_{\mathbf{a}\to\mathbf{e}} \leq exp\{-\frac{|\boldsymbol{\eta}|^4}{\sigma^2}\frac{E_s}{4N_0}\}.$$

Theorem 2 *Suppose Viterbi Algorithm is used for decoding for equation (3), then the pair-wise symbol error probability of transmitting* $\mathbf{a} = (a_1, a_2, \cdots, a_m)$ *and decoding in favor of* $\mathbf{e} = (e_1, e_2, \cdots, e_m)$ *is bounded by*

$$P_{\mathbf{a}\to\mathbf{e}} \leq exp\{-\frac{|\boldsymbol{\eta}|^4}{\sigma^2}\frac{E_s}{4N_0}\}, \tag{16}$$

where, σ^2 *is given in* (14).

Hence, a good approach to minimizing the error probability is therefore to maximize

$$\min_{\mathbf{B},\mathbf{C}} \frac{|\boldsymbol{\eta}|^4}{\sigma^2}.$$

IV. BI-TRUNCATION

In this section, we focus on bi-truncation. We will first establish criteria for bi-truncation and then develop two approaches.

A. BI-TRUNCATION CRITERIA

For simplicity, we write the truncation matrix in (2) for bi-truncation as in the following,

$$\mathbf{B} = \begin{pmatrix} d_1 & c_1 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & c_2 & \cdots & 0 & 0 \\ 0 & 0 & d_3 & \ddots & 0 & 0 \\ \cdots & \cdots & \cdots & \ddots & \ddots & \cdots \\ 0 & 0 & 0 & \cdots & d_{m-1} & c_{m-1} \\ c_m & 0 & 0 & \cdots & 0 & d_m \end{pmatrix} \tag{17}$$

where $d_i \neq 0$ and $c_i \neq 0$ for all $1 \leq i \leq m$.

At high SNR, when $\mathbf{a}$ is transmitted, and the receiver decodes to $\mathbf{e}$, usually at most one bit is detected wrong, that is, $\mathbf{a}$ and $\mathbf{e}$ differ in at most one bit. The bound in (16) allows us to derive an upper bound on the bit error probability, based on which bi-truncation design criteria will be derived. Suppose $\mathbf{a}$ and $\mathbf{e}$ differ in only one bit, and assume $a_l \neq e_l$ for some l ($1 \leq l \leq m$). Then $\boldsymbol{\eta}$ has only two nonzero components $\eta_{l-1} = c_{l-1}(a_l - e_l)$ and $\eta_l = d_l(a_l - e_l)$. Thus $|\boldsymbol{\eta}|^4 = (|d_l|^2 + |c_{l-1}|^2)^2|a_l - e_l|^4$.

Throughout the rest of the paper, we adopt the following subscript operation

$$\begin{aligned} i+1 &:= \delta(i-m) + (i+1) \bmod (m+1) \\ i-1 &:= m\delta(i-1) + (i-1) \bmod m \end{aligned}$$

where $\delta(\cdot)$ is the Deta function, i.e.,

$$\delta(x) = \begin{cases} 1 & \text{if } x = 0; \\ 0 & \text{otherwise.} \end{cases}$$

Now

$$\begin{aligned} \sigma^2 = &\ \textstyle\sum_{k=1}^m (Re(\eta_l \mathbf{C}^*_{l,k}) + Re(\eta_{l-1}\mathbf{C}^*_{l-1,k}))^2 \\ &+ \textstyle\sum_{k=1}^m (Im(\eta_l \mathbf{C}^*_{l,k}) + Im(\eta_{l-1}\mathbf{C}^*_{l-1,k}))^2 \\ = &\ |\eta_l|^2(\mathbf{CC}^*)_{l,l} + |\eta_{l-1}|^2(\mathbf{CC}^*)_{l-1,l-1} \\ &+ 2Re(\eta_l\eta^*_{l-1}(\mathbf{CC}^*)_{l-1,l}). \end{aligned} \tag{18}$$

Recall that $\mathbf{CC}^* = \mathbf{BXB}^* + \mathbf{C}_2\mathbf{C}_2^*$. By Cauchy-Schwartz inequality,

$$((\mathbf{C}_2\mathbf{C}_2^*)_{l,l}(\mathbf{C}_2\mathbf{C}_2^*)_{l-1,l-1})^{1/2} \geq |(\mathbf{C}_2\mathbf{C}_2^*)_{l-1,l}|.$$

With this inequality, we can show

$$\begin{aligned} \sigma^2 \geq &\ |\eta_l|^2(\mathbf{BXB}^*)_{l,l} + |\eta_{l-1}|^2(\mathbf{BXB}^*)_{l-1,l-1} \\ &+ 2Re(\eta_l\eta^*_{l-1}(\mathbf{BXB}^*)_{l-1,l}), \end{aligned}$$

with equality if and only if $\mathbf{C}_2$ is a zero matrix—this is consistent with the maximum SNR approach. Hence, throughout the rest of the paper, we set $\mathbf{C}_2 = \mathbf{0}$. As a result, $\mathbf{CC}^* = \mathbf{BXB}^*$.

Thus, in the case when $\mathbf{a}$ and $\mathbf{e}$ differ only in the lth bit, the pairwise symbol error probability becomes the lth bit error probability and is bounded by

$$P_l \leq exp\{-\frac{(|d_l|^2 + |c_{l-1}|^2)^2}{\sigma_l^2}\frac{E_s}{4N_0}|a_l - e_l|^2\}, \tag{19}$$

where

$$\begin{aligned} \sigma_l^2 = &\ |\eta_l|^2(\mathbf{BXB}^*)_{l,l} + |\eta_{l-1}|^2(\mathbf{BXB}^*)_{l-1,l-1} \\ &+ 2Re(\eta_l\eta^*_{l-1}(\mathbf{BXB}^*)_{l-1,l}) \\ = &\ |d_l|^2(\mathbf{BXB}^*)_{l,l} + |c_{l-1}|^2(\mathbf{BXB}^*)_{l-1,l-1} \\ &+ 2Re(d_l c^*_{l-1}(\mathbf{BXB}^*)_{l-1,l}). \end{aligned} \tag{20}$$

The probability that any single bit decision is incorrect will be dominated by the largest bit error probability, or equivalently by the min $_{1\leq l\leq m}$ $\frac{(|d_l|^2 + |c_{l-1}|^2)^2}{\sigma_l^2}\}$. Hence, we conclude that the optimal choice of $\mathbf{B}$ should maximize

$$\min_{1\leq l\leq m} \frac{(|d_l|^2 + |c_{l-1}|^2)^2}{\sigma_l^2}. \tag{21}$$

It is non-trivial to find an optimal solution $\mathbf{B}$ for (21). However, from the upper bound on the error probability, we see that since tr($\mathbf{BB}^*$) is fixed, if for some i ($1 \leq i \leq m$), $|d_i|$ or $|c_i|$ is too small, then there must exist some other $k \neq i$, such that $|d_k|$ or $|c_k|$ is large. As a result, (21) could not be maximized. Therefore, $|d_i|$, $|c_i|$, $1 \leq i \leq m$, should be approximately equal. Therefore, we propose the following criteria for designing a bi-truncation matrix $\mathbf{B}$:

- $|d_i|$, $|c_i|$, $1 \leq i \leq m$, *are approximately equal.*
- *Minimize tr*($\mathbf{BXB}^*$) *while fixing tr*($\mathbf{BB}^*$).

A good strategy is to choose $\mathbf{B}$ so as to balance these two criteria. We formalize the bi-truncation problem as follows: Find bi-diagonal matrix $\mathbf{B}$ to minimize

$$\begin{aligned}\text{tr}(\mathbf{BXB}^*) = \textstyle\sum_{i=1}^m(&|d_i|^2 x_{i,i} + |c_i|^2 x_{i+1,i+1}\\ &+2Re(d_i c_i^* x_{i,i+1}))\end{aligned} \tag{22}$$

while tr($\mathbf{BB}^*$) is fixed and $|d_i|$, $|c_i|$, $1 \leq i \leq m$, are approximately equal.

For technical convenience, we set tr($\mathbf{BB}^*$) $= m$, where m is the number of transmit antennas. Note that once $\mathbf{B}$ is selected, $\mathbf{C}$ is determined by (5) with $\mathbf{C}_2 = \mathbf{0}$.

B. BI-TRUNCATION APPROACHES

In this section, two bi-truncation approaches are introduced. It seems unrealistic to find a closed-form solution for $\mathbf{B}$, so we attempt to find a numerical solution. Since Re$(d_i c_i^* x_{i,j}) \geq -|d_i||c_i||x_{i,j}|$,

$$\begin{aligned}\text{tr}(\mathbf{BXB}^*) \geq \textstyle\sum_{i=1}^m(&|d_i|^2 x_{i,i} + |c_i|^2 x_{i+1,i+1}\\ &-2|d_i||c_i||x_{i,i+1}|).\end{aligned} \tag{23}$$

The equality in (23) is achieved if and only if

$$\begin{gathered} x_{i,i+1} = 0 \\ \text{or} \\ d_i c_i^* = \text{a negative scalar of } x^*_{i,i+1}. \end{gathered} \tag{24}$$

Hence, the optimal phase relationship between d_i, c_i and $x_{i,i+1}$ is well determined for all i, and we only need to consider $|d_i|$ and $|c_i|$, $1 \leq i \leq m$, the magnitudes of the entries of $\mathbf{B}$. We shall explore two bi-truncation approaches.

1) Approach I: In this subsection, we explore one approach by maximizing SNR. The objective is to choose $\mathbf{B}$ that minimizes

$$\begin{aligned}\text{tr}(\mathbf{BXB}^*) = \textstyle\sum_{i=1}^m(&|d_i|^2 x_{i,i} + |c_i|^2 x_{i+1,i+1}\\ &-2|d_i||c_i||x_{i,i+1}|)\end{aligned} \tag{25}$$

subject to the constraint $\sum_{i=1}^m |d_i|^2 + |c_i|^2 = m$.

Using Lagrange multiplier, it can be easily shown that there is no interior critical point, which implies that the minimum is achieved on the boundary where at least one of $|d_i|$'s or $|c_i|$'s is 0. Hence, we take an alternative approach.

For each i, set $|d_i| = L_i\text{sin}\theta_i$, $|c_i| = L_i\text{cos}\theta_i$, $\theta_i \in (0, \frac{\pi}{2})$. Then $|d_i|^2 + |c_i|^2 = L_i^2$, and $\sum_{i=1}^m |L_i|^2 = m$. We require that $L_i^2 \geq \epsilon$ ($i = 1, 2, \cdots, m$), where $0 < \epsilon \leq 1$. tr($\mathbf{BXB}^*$) can be computed as follows.

$$\begin{aligned}&(\mathbf{BXB}^*)_{i,i} = |d_i|^2 x_{i,i} + |c_i|^2 x_{i+1,i+1} - 2|d_i||c_i||x_{i,i+1}|\\ &= L_i^2(x_{i,i}\text{sin}^2\theta_i + x_{i+1,i+1}\text{cos}^2\theta_i - |x_{i,i+1}|\text{sin}2\theta_i)\\ &= L_i^2(\frac{x_{i,i} + x_{i+1,i+1}}{2} - k_i\text{cos}(2\theta_i - \phi_i))\\ &\geq L_i^2(\frac{x_{i,i} + x_{i+1,i+1}}{2} - k_i),\end{aligned}$$

with equality if and only if $\theta_i = \dfrac{\phi_i}{2}$, where

$$k_i = ((\frac{x_{i,i} - x_{i+1,i+1}}{2})^2 + |x_{i,i+1}|^2)^{1/2}$$

and

$$\phi_i = \text{cos}^{-1}(\frac{x_{i,i} - x_{i+1,i+1}}{2k_i}).$$

Set $S_i = \dfrac{x_{i,i} + x_{i+1,i+1}}{2} - k_i$, $S_{i_0} = min_{\ 1 \leq l \leq m}\ S_i$. Then by the assumptions that $L_i^2 \geq \epsilon$ and $\sum_{i=1}^m L_i^2 = m$, and since $S_{i_0} \leq S_i$ for $i \neq i_0$, we have

$$\begin{aligned}&\text{tr}(\mathbf{BXB}^*) = \textstyle\sum_{i=1}^m L_i^2 S_i\\ &\geq (\textstyle\sum_{i \neq i_0} \epsilon S_i) + (m - (m-1)\epsilon)S_{i_0}.\end{aligned} \tag{26}$$

The equality holds in (26) if and only if $L_{i_0}^2 = m - (m-1)\epsilon$ and $L_i^2 = \epsilon$ for all $i \neq i_0$. Therefore, tr($\mathbf{BXB}^*$) is minimized by taking

$$\begin{gathered}|d_i| = \begin{cases} \sqrt{\epsilon}\ \text{sin}\dfrac{\phi_i}{2} & \text{if } i \neq i_0; \\ \sqrt{m-(m-1)\epsilon}\ \text{sin}\dfrac{\phi_i}{2} & \text{if } i = i_0; \end{cases}\\ \\ |c_i| = \begin{cases} \sqrt{\epsilon}\ \text{cos}\dfrac{\phi_i}{2} & i \neq i_0; \\ \sqrt{m-(m-1)\epsilon}\ \text{cos}\dfrac{\phi_i}{2} & i = i_0. \end{cases}\end{gathered} \tag{27}$$

Combining (24) and (27), we minimize tr$(\mathbf{BXB}^*)$ by setting

$$d_i = \begin{cases} |d_i| & \text{if } x_{i,i+1} = 0; \\ |d_i| \dfrac{x^*_{i,i+1}}{|x_{i,i+1}|} & \text{if } x_{i,i+1} \neq 0; \end{cases} \tag{28}$$

$$c_i = -|c_i|.$$

The bi-truncation matrix derived in (28) is optimal in the sense of maximum average SNR.

2) Approach II: In approach I, the minimum of tr$(\mathbf{BXB}^*)$ has the asymptotic property:

$$\text{tr}(\mathbf{BXB}^*) \to mS_{i_0} \quad \text{as} \quad \epsilon \to 0.$$

However, SNR is not the unique factor that affects the performance of a scheme. If ϵ is small, then all $|d_i|$'s and $|c_i|$'s $(i \neq i_0)$ are very small, which is equivalent to deep fading in $m-1$ subchannels. Therefore, there is a tradeoff between the SNR and the fading. From the error probability derivation (21), such truncation $\mathbf{B}$ for small ϵ needs not provide the best performance even though it maximizes the SNR.

Next, we consider both SNR and the fading in search of a bi-truncation matrix $\mathbf{B}$. It turns out that this matrix $\mathbf{B}$ is very easy to compute and implement. Notice that (25) may be written as

$$\text{tr}(\mathbf{BXB}^*) = \textstyle\sum_{i=1}^m (|d_i|\sqrt{x_{i,i}} - |c_i|\sqrt{x_{i+1,i+1}})^2$$

$$+2|d_i||c_i|(\sqrt{x_{i,i}x_{i+1,i+1}} - |x_{i,i+1}|).$$

Let

$$|d_i| = \frac{\sqrt{x_{i+1,i+1}}}{s_i} \quad , \quad |c_i| = \frac{\sqrt{x_{i,i}}}{s_i}, \tag{29}$$

where $s_i = \sqrt{x_{i,i} + x_{i+1,i+1}}$. From $\mathbf{X} = (\mathbf{V}^*\mathbf{D}^{-1})(\mathbf{V}^*\mathbf{D}^{-1})^*$, we see that no $|d_i|$ or $|c_i|$ for all $1 \leq i \leq m$ will result in deep fading, and (25) is close to be minimized. Let $\mathbf{B}$ be the bi-truncation matrix with entries

$$d_i = \begin{cases} \dfrac{\sqrt{x_{i+1,i+1}}}{s_i} & \text{if } x_{i,i+1} = 0; \\ \dfrac{\sqrt{x_{i+1,i+1}}}{s_i} \cdot \dfrac{x^*_{i,i+1}}{|x_{i,i+1}|} & \text{if } x_{i,i+1} \neq 0; \end{cases} \tag{30}$$

$$c_i = -\frac{\sqrt{x_{i,i}}}{s_i}.$$

The simulation results show that this bi-truncation well balances two design criteria we proposed in Section IV.

C. GROUPING OF BI-TRUNCATION

Up to this point, we have assumed that the bi-truncation matrix $\mathbf{B}$ is a bi-diagonal matrix, which allows the modified Viterbi detector to detect the symbols in the natural grouping $(a_1, a_2), (a_2, a_3), \cdots, (a_m, a_1)$ of $a_1, a_2, \cdots, a_m$. Similar to the sorted MMSE-DFD, this is not the only possible grouping. Note that each grouping forms an array, under rotation and/or reflection, the array still corresponds to an equivalent grouping. This is equivalent to a Dihedral Group $\boldsymbol{D_{2m}}$ acting on the set comprised of all these arrays. The order of the Dihedral Group is $2m$. By applying Burnside's theorem [12], there are

$$\frac{m!}{2m} = \frac{(m-1)!}{2}$$

nonequivalent groupings.

Each grouping determines a bi-truncation matrix $\mathbf{B}$, which is not necessarily bi-diagonal. Suppose $(a_1, a_{i_2}), (a_{i_2}, a_{i_3}), \cdots, (a_{i_m}, a_1)$ is a grouping, which corresponds to $\mathbf{a}' = (a_1, a_{i_2}, a_{i_3}, \cdots, a_{i_m})$, then there is a permutation matrix $\mathbf{P}$ such that $\mathbf{a} = \mathbf{P}^T\mathbf{a}'$, where $\mathbf{a} = (a_1, a_2, a_3, \cdots, a_m)$. Substituting $\mathbf{a} = \mathbf{P}^T\mathbf{a}'$ into the truncation channel model (3) yields

$$\tilde{\mathbf{r}} = \mathbf{BP}^T\mathbf{a}' + \tilde{\mathbf{z}} = \mathbf{B}'\mathbf{a}' + \tilde{\mathbf{z}}, \tag{31}$$

where $\mathbf{B}'$ is a bi-diagonal matrix. Making use of bi-truncation II, $\mathbf{B}'$ can be computed, and then $\mathbf{B} = \mathbf{B}'\mathbf{P}$ is the bi-truncation matrix corresponding to the grouping $(a_1, a_{i_2}), (a_{i_2}, a_{i_3}), \cdots, (a_{i_m}, a_1)$. The best grouping corresponds to the bi-truncation matrix $\mathbf{B}$ which minimizes

$$\text{tr}(\mathbf{B}'\mathbf{XB}'^*) = \text{tr}(\mathbf{BX}'\mathbf{B}^*),$$

where $\mathbf{X}' = \mathbf{P}^T\mathbf{XP}$. For example, there are 3 nonequivalent groupings when $m = 4$, so there are 3 nonequivalent bi-truncation matrices:

$$\mathbf{B}_1 = \begin{pmatrix} d_1 & c_1 & 0 & 0 \\ 0 & d_2 & c_2 & 0 \\ 0 & 0 & d_3 & c_3 \\ c_4 & 0 & 0 & d_4 \end{pmatrix},$$

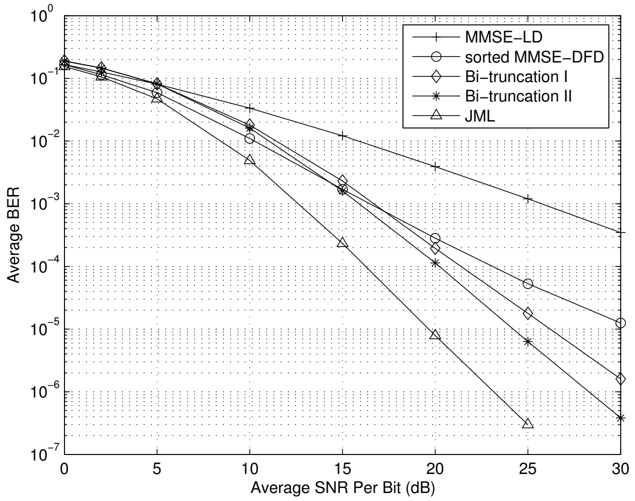

Fig. 2. Comparison performance of uncoded 4-QAM codes with 3 Tx and 3 Rx.

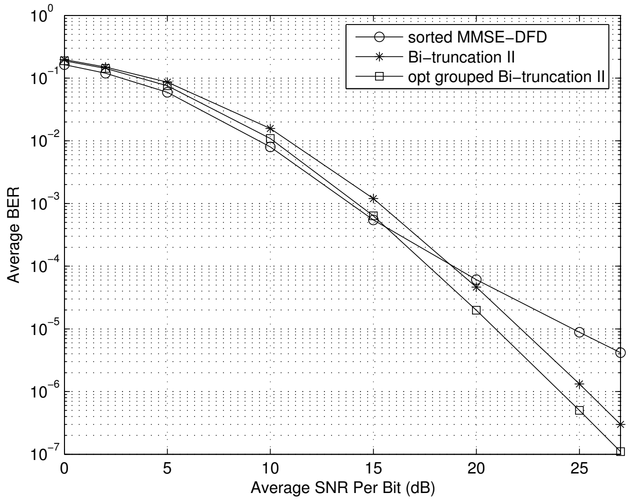

Fig. 3. Comparison performance of uncoded 4-QAM codes with 4 Tx and 4 Rx.

$$\mathbf{B}_2 = \begin{pmatrix} d_1 & 0 & c_1 & 0 \\ 0 & c_2 & d_2 & 0 \\ 0 & d_3 & 0 & c_3 \\ c_4 & 0 & 0 & d_4 \end{pmatrix},$$

$$\mathbf{B}_3 = \begin{pmatrix} d_1 & 0 & c_1 & 0 \\ 0 & 0 & d_2 & c_2 \\ 0 & c_3 & 0 & d_3 \\ c_4 & d_4 & 0 & 0 \end{pmatrix}.$$

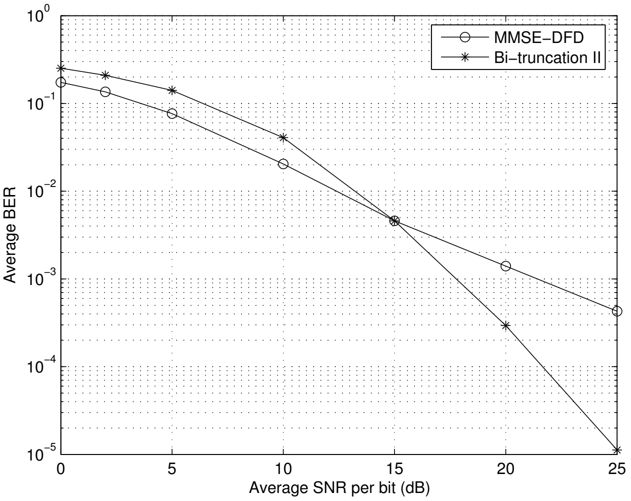

Fig. 4. Comparison performance of uncoded 4-QAM codes with 8 Tx and 8 Rx.

V. SIMULATION RESULTS

In this section, we compare the performance of bi-truncation with MMSE-LD and MMSE-DFD for memoryless MIMO systems. Instead of specializing to a particular channel matrix, we will average the performance over one million to ten million randomly generated Rayleigh fading channels. All transmit signals are chosen uniformly and independently from 4-QAM alphabet. The performance of bi-truncation detection is obtained by modified Viterbi Algorithm.

In Figure 2, a system with 3 transmit antennas and 3 receive antennas is simulated. In this system, both bi-truncation detection approaches outperform MMSE-LD and sorted MMSE-DFD at high SNR. Also, bi-truncation II outperforms bi-truncation I ($\epsilon = 1$), the main reason is that Bi-truncation I results in deep fading sometimes, while Bi-truncation II provides a good balance between SNR and fading. A 4×4 system is simulated in Figure 3, the simulation results indicate that Bi-truncation II and optimal grouped Bi-truncation II both outperform sorted MMSE-DFD at high SNR. In Figure 4 we simulate an 8×8 system, we notice that Bi-truncation II provides much better performance than MMSE-DFD.

VI. CONCLUSIONS

In this paper, we propose a channel truncation based detector for low complexity MIMO signal detection, and give theoretical analysis into the design and performance. The bi-truncation is studied in detail. Two bi-truncation approaches are presented, and the simulation results indicate that bi-truncation detector outperforms MMSE-LD and sorted MMSE-DFD when a system is equipped with a small number of transmit antennas. Especially, the average *bit error probability* (BER) by bi-truncation detection decreases dramatically as SNR increases. When implementing SD, the major issues are choosing the initial radius and the order in which the inputs are detected. The advantage of bi-truncation detection over SD is therefore the existence of simple modified Viterti Algorithm decoding.

References

[1] S. Verdú, *Multiuser Detection*, Cambridge University Press, 1998.

[2] A. Duel-Hallen, ”Decorrelating decision-feedback multiuser detector for synchronous code-division multiple access channel,” *IEEE Trans. Commun.*, vol 41, N0.2, pp. 285-290, Feb. 1993.

[3] G. Foschini, G. Golden, R. Valenzuela, P., Wolniansky, ”Simplified processing for wireless communication at high spectral efficiency,” *IEEE J. Select. Areas Commun.*, vol 17, No. 11, pp. 1841-1852, Nov. 1999.

[4] O. Damen, A. Chkeif, J. Belfiore, ”Lattice code decoder for space-time codes,” *IEEE Commun. Letters*, vol. 4, no. 5, pp. 161-163, May 2000.

[5] Ami Wiesel, Xavier mestre, Alba Pages, and Javier R. Fonollosa, ”Efficient implementation of sphere demodulation,” *IEEE SPAWC*, 2003.

[6] D. D. Falconer, and F. R. Magee, JR., ”Adaptive channel memory truncation for maximum likelihood sequence estimation,” *The Bell System Technical J.*, vol. 52, No. 9, pp. 1541-1562, Nov. 1973.

[7] Jeremiah F. Hayes, ”The Viterbi algorithm applied to digital data transmission,” *IEEE Commun. Mag.*, pp. 26-32, May 2002.

[8] Stephen B. Wicker, Error Control Systems for Digital Communication and Storage, Prentice Hall 1995.

[9] D. Divsalar and M.k. Simon, ”Trellis coded modulation for 4800 to 9600 bps transmission over a fading satellite channel,” *IEEE J. Select. Areas Commun.*, vol. SAC-5, pp. 162-175, Feb. 1987.

[10] J. K. Cavers and P. Ho, ”Analysis of the error performance of trellis coded modulation in Rayleigh fading channels,” *IEEE Trans. Commun.*, vol 40, pp. 74-83, Jan. 1992.

[11] Noga Alon, Joel H. Spencer, The Probabilistics Method, Second Edition, John Wiley and Sons Inc, 2000.

[12] Richard A. Brualdi, Introductory Combinatorics, Third Edition, Prentice Hall 1999.