

# Improvement in Accuracy of Signal Detection Aided by Element-Based Lattice Reduction for Massive MIMO Wireless Communication Systems

Tatsuki FUKUDA

## Abstract

A Massive MIMO wireless communication system is an important factor for the next generation of mobile communication, but the computational complexity for signal detection is a problem due to many antennas the systems employed. The lattice-reduction-aided linear detection method is one of the solution, and the element-based lattice reduction is a lattice reduction technique that requires a practical computational complexity. We found that there still be the wasted process in the element-based lattice reduction, so we propose the improved method. In some experiments, we show that our method can reduce 63.3% of the number of iterations in detecting process compared to original via the number of antennas in exchange for just 35.4% of the bit-error-ratio performance.

*Key words: Massive MIMO, Element-based lattice reduction, Signal detection*

## 1. Introduction

Massive MIMO (Multiple-Input Multiple-Output) wireless communication systems are receiving a lot of attention as the important technique for the 5th generation of mobile communication systems. Massive MIMO systems employ many antennas for transmit and receive, so it can meet the requirements of consumers that the higher data-rate or more reliable communication than ever. It, however, has some problems such as a signal detection. The best performance of bit error ratio (BER) is obtained by the maximum likelihood detection (MLD) method<sup>[1]</sup>. With the MLD method, the detector calculates the Euclidean distance between the received signal and all of the signal candidates and choose the nearest signal point in constellation as the estimated signal. Instead of the best BER performance, the computational complexity in MLD exponentially increase according to the number of transmit antennas<sup>[2]</sup>. In view of the complexity, the best method is a kind of linear detection (LD) such as zero forcing (ZF) methods and minimum mean square error (MMSE) methods. These methods have linear complexity, but the BER performances of them are inferior to that of MLD. The massive MIMO wireless communication systems have a lot of transmit antennas. That's mean that the LD is the best way for the system because the complexity of MLD or some kinds of improved MLD increase with the increase in the number of antennas, and some kinds of improved MLD are not practical in view of the computational complexity.

Recently, the Lenstra-Lenstra-Lovasz (LLL) algorithm is so attractive to researchers of massive MIMO. The LLL is one of lattice reduction (LR) algorithms, and the LLL can obtain the nearly orthogonal lattice basis in polynomial time<sup>[3]</sup>. The LLL algorithm is used as a pre-process of LD, and the orthogonality of each signals transmitted at the same timeslot are enlarged in order to detect them easily. The LLL algorithm can obtain the orthogonality in polynomial time, but the computational complexity of the LLL is not small enough. In order to solve the problem, element-based lattice reduction (ELR) algorithm is proposed<sup>[4]</sup>. ELR algorithm tries to minimize the diagonal elements in noise covariance matrix, and it requires lower

computational complexity than The LLL. In this paper, we indicate the waste point of ELR algorithm in view of the computational complexity, and propose the solution for it.

## 2. System Model and ZF Detection Method

The system model is shown in Fig. 1. The signals are transmitted from  $N_T$  antennas and reached  $N_R$  antennas at receiver. The signals pass through the additive white Gaussian noise (AWGN) and quasi-static Rayleigh fading channel.

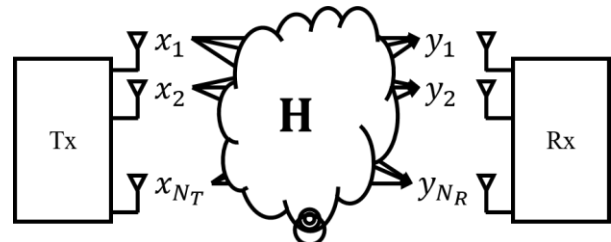


Fig. 1 System model of massive MIMO.

In this paper, we consider the block transmission model shown as

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{N}, \quad (1)$$

where  $\mathbf{Y} = [y_1 \ y_2 \ \dots \ y_{N_R}]^T$  and  $\mathbf{X} = [x_1 \ x_2 \ \dots \ x_{N_T}]^T$  denote the received signal vector and the transmitted signal vector, respectively.  $\mathbf{N} = [n_1 \ n_2 \ \dots \ n_{N_R}]^T$  denotes the noise vector with zero mean and covariance matrix  $\sigma_w^2 \mathbf{I}_{N_R}$ , where  $\mathbf{I}_{N_R}$  is  $N_R \times N_R$  identity matrix.  $\mathbf{H}$  denotes an  $N_R \times N_T$  complex channel matrix. The transmitted signal vector is guessed from the received signal vector as the estimated signal vector  $\mathbf{X}' = [x'_1 \ x'_2 \ \dots \ x'_{N_T}]^T$  in detection process.

The easiest way to guess the transmitted signal is to multiply  $\mathbf{H}^{-1}$  to  $\mathbf{Y}$ , but  $\mathbf{H}$  is not always the squared matrix, so the Moore-Penrose pseudo inverse of  $\mathbf{H}$  is used in ZF method. Thus, we obtain the estimated signal vector  $\mathbf{X}'_{ZF}$  shown in Eq. 2<sup>[5]</sup>.

$$\mathbf{X}'_{ZF} = \mathbf{Q}(\mathbf{H}^\dagger \mathbf{Y}) = \mathbf{Q}((\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \mathbf{Y}), \quad (2)$$

where  $\mathbf{H}^\dagger$  denotes the Moore-Penrose pseudo inverse of  $\mathbf{H}$ ,  $\mathbf{H}^H$  is Hermitian matrix of  $\mathbf{H}$ , and  $\mathbf{Q}(\cdot)$  is the symbol-wise quantizer to the constellation set.

### 3. Lattice-Reduction-Aided Detection

MLD realize the best BER performance, but its complexity is so high. ZF is one of the fastest detection method, but its accuracy is not so good. As you can guess, making the accuracy of ZF higher or making complexity of MLD lower are the solution. The former is the basic policy of the lattice-reduction-aided linear detection.

A lattice is defined as

$$\mathcal{L} = \{\sum_{i=1}^N \mathbf{a}_i b_i \mid \mathbf{a}_i \in \mathbb{Z}[j]\}, \quad (3)$$

where  $\mathbb{Z}[j]$  denotes the Gaussian integer ring whose elements have a form  $\mathbb{Z} + j\mathbb{Z}$ ,  $b_i (i = 1, \dots, N)$  denotes the basis vectors of lattice  $\mathcal{L}$ . The real and imaginary parts of  $x_i$  are one of  $\{2m + 1 - \sqrt{M}, m = 0, 1, \dots, \sqrt{M} - 1\}$  if  $M$ -QAM is employed, so  $x_i \in \mathbb{Z}[j]$ . It follows that  $\mathbf{H}\mathbf{X} \in \mathcal{L}$ , where the basis is the columns of  $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_{N_T}]$ .

In order to find more ‘‘orthogonal’’ basis, LR algorithm reduce the lattice basis. The process is equivalent to find a unimodular matrix  $\mathbf{T}$  such that  $\tilde{\mathbf{H}} = \mathbf{H}\mathbf{T}$  make the same lattice as  $\mathbf{H}$ <sup>[6]</sup>. If the number of transmit antennas is two, the Gaussian reduction algorithm<sup>[9]</sup> is optimal, and various LR algorithms are proposed for the system employed more antennas than 2<sup>[5]</sup>. Anyway, we get the Eq. 4 after multiply the  $\tilde{\mathbf{H}}^\dagger$  to  $\mathbf{Y}$ .

$$\mathbf{W}' = \tilde{\mathbf{H}}^\dagger \mathbf{Y} = \mathbf{T}^{-1} \mathbf{X} + \tilde{\mathbf{H}}^\dagger \mathbf{N} \equiv \mathbf{Z} + \mathbf{N}'. \quad (4)$$

As the elements of  $\mathbf{X}$  are parts of the  $M$ -QAM set, the real and imaginary parts of the element of  $\mathbf{b} = \mathbf{T}^{-1} \{\mathbf{X} - (1+j)\mathbf{1}_{N_T \times 1}\}/2$ , where  $\mathbf{1}_{k \times m}$  is the  $k \times m$  matrix whose elements are all 1, are in consecutive integer sets. Thus we obtain the estimate of  $\mathbf{Z}$ <sup>[10, 11]</sup>,

$$\tilde{\mathbf{Z}} = 2\tilde{\mathbf{b}} + (1+j)\mathbf{T}^{-1}\mathbf{1}_{N_T \times 1}, \quad (5)$$

where  $\tilde{\mathbf{b}} = \lfloor (\mathbf{W} - (1+j)\mathbf{T}^{-1}\mathbf{1}_{N_T \times 1})/2 \rfloor$  and  $\lfloor \cdot \rfloor$  denotes a rounding function. Finally, we obtain estimated signal vector  $\mathbf{X}'_{LRZF}$  by ZF as Eq. 6.

$$\mathbf{X}'_{LRZF} = \mathbf{Q}(\mathbf{T}\tilde{\mathbf{Z}}) = \mathbf{Q}\left(\mathbf{X} + 2\mathbf{T}\left[\frac{1}{2}\tilde{\mathbf{H}}^\dagger \mathbf{N}\right]\right), \quad (6)$$

### 4. ELR Algorithm and Problem Statement

The ELR is a LR technique and it can find the reduced lattice in view of pair-wise error ratio (PEP). The ELR algorithm obtain  $\tilde{\mathbf{H}}$  and  $\mathbf{T}$  from  $\mathbf{C} = (\mathbf{H}^H \mathbf{H})^{-1}$  which is a scaled covariance matrix of

the noise after equalization.

First, ELR choose a pair of indices  $(i, k)$ , determine  $\lambda_{i,k} \in \mathbb{Z}[j]$ , and update the  $k$ -th column of matrix  $\mathbf{T}' = (\mathbf{T}^{-1})^H$  as

$$\mathbf{t}'_k = \mathbf{t}'_k + \lambda_{i,k} \mathbf{t}'_i, \quad (7)$$

where  $\mathbf{t}'_k$  is the  $k$ -th column of  $\mathbf{T}'$ . The  $k$ -th column and  $k$ -th row of  $\tilde{\mathbf{C}} = (\tilde{\mathbf{H}}^H \tilde{\mathbf{H}})^{-1}$  are also updated as follows,

$$\tilde{\mathbf{c}}_k = \tilde{\mathbf{c}}_k + \lambda_{i,k} \tilde{\mathbf{c}}_i \quad (8)$$

$$\tilde{\mathbf{c}}^{(k)} = \tilde{\mathbf{c}}^{(k)} + \lambda_{i,k}^* \tilde{\mathbf{c}}^{(i)}, \quad (9)$$

where,  $\tilde{\mathbf{c}}_k$  and  $\tilde{\mathbf{c}}^{(k)}$  are  $k$ -th column and  $k$ -th row of  $\tilde{\mathbf{C}}$ , respectively, and superscript \* denotes the complex conjugate. These updates make  $\tilde{\mathbf{H}}$  be updated as follows,

$$\tilde{\mathbf{h}}_i = \tilde{\mathbf{h}}_i - \lambda_{i,k} \tilde{\mathbf{h}}_k, \quad (10)$$

where,  $\tilde{\mathbf{h}}_i$  is  $i$ -th column of  $\tilde{\mathbf{H}}$ . Thu,  $\tilde{\mathbf{c}}_{k,k}$ , which is the  $(k, k)$  element of  $\tilde{\mathbf{C}}$ , is updated as

$$\tilde{\mathbf{c}}_{k,k} = \tilde{\mathbf{c}}_{k,k} + |\lambda_{i,k}|^2 \tilde{\mathbf{c}}_{i,i} + \lambda_{i,k}^* \tilde{\mathbf{c}}_{i,k} + \lambda_{i,k} \tilde{\mathbf{c}}_{k,i}. \quad (11)$$

Note that the diagonal elements except for  $\tilde{\mathbf{c}}_{k,k}$  is not updated. The amount of decrease of  $\tilde{\mathbf{c}}_{k,k}$  is  $\Delta_{i,k}$  represented as Eq. 12, and  $\tilde{\mathbf{c}}_{k,k}$  is minimized when  $\lambda_{i,k}$  satisfies Eq. 13<sup>[5]</sup>.

$$\Delta_{i,k} = -|\lambda_{i,k}|^2 \tilde{\mathbf{c}}_{i,i} - \lambda_{i,k}^* \tilde{\mathbf{c}}_{i,k} - \lambda_{i,k} \tilde{\mathbf{c}}_{k,i}. \quad (12)$$

$$\lambda_{i,k} = -\left\lfloor \frac{\tilde{\mathbf{c}}_{i,k}}{\tilde{\mathbf{c}}_{i,i}} \right\rfloor. \quad (13)$$

ELR algorithm terminates the process when  $\lambda_{i,k}$  becomes zero. The ELR algorithm is summarized in Table 1.

**Table 1** Element-based lattice reduction algorithm

Input: $\mathbf{H}$	Output: $\tilde{\mathbf{H}}, \mathbf{T}$
(1)	$\tilde{\mathbf{C}} = (\mathbf{H}^H \mathbf{H})^{-1}, \mathbf{T}' = \mathbf{I}_N$
(2)	<b>Do</b>
(3)	$\lambda_{i,k} = -\lfloor \tilde{\mathbf{c}}_{i,k} / \tilde{\mathbf{c}}_{i,i} \rfloor$
(4)	<b>If all</b> $\lambda_{i,k} = 0, \forall i \neq k$ , <b>goto</b> (11)
(5)	<b>Find the largest reducible</b> $\tilde{\mathbf{c}}_{k,k}$
(6)	<b>Choose</b> $i = \arg \max_{i=1, i \neq k}^{N_T} \Delta_{i,k}$
(7)	$\mathbf{t}'_k = \mathbf{t}'_k + \lambda_{i,k} \mathbf{t}'_i$
(8)	$\tilde{\mathbf{c}}_k = \tilde{\mathbf{c}}_k + \lambda_{i,k} \tilde{\mathbf{c}}_i$
(9)	$\tilde{\mathbf{c}}^{(k)} = \tilde{\mathbf{c}}^{(k)} + \lambda_{i,k}^* \tilde{\mathbf{c}}^{(i)}$
(10)	<b>While(true)</b>
(11)	$\mathbf{T} = (\mathbf{T}'^{-1})^H, \tilde{\mathbf{H}} = \mathbf{H}\mathbf{T}$

By the way, the PEP of LR-aided ZF detection is calculated [5]

as

$$P(\mathbf{z}_i \rightarrow \tilde{\mathbf{z}}_i | \mathbf{H}) = A \left( \frac{\sqrt{|e_{z_i}|^2}}{\sqrt{2\sigma_w^2 \tilde{\mathbf{c}}_{i,i}}} \right), \quad (14)$$

where  $z_i$  and  $\tilde{z}_i$  denote  $i$ -th element of  $\mathbf{Z}$  in Eq. 4 and  $\tilde{\mathbf{Z}}$  in Eq. 5, respectively,  $e_{z_i} = z_i - \tilde{z}_i$ , and

$$A(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{t^2}{2}\right) dt. \quad (15)$$

As you see, the PEP is determined by  $\tilde{c}_{i,i}$  especially at high signal to noise ratio (SNR), so the LR algorithm reduce the diagonal elements of  $\tilde{\mathbf{C}}$ . ELR also tries to do so, but the PEP is influenced by  $e_{z_i}$  and  $\sigma_w^2$ . It means that the diagonal elements of  $\tilde{\mathbf{C}}$  have to be small enough but not need to be minimized. ELR algorithm minimize them, so there is unnecessary process in it.

### 5. Improvement of ELR

As we said, diagonal elements have to be just small enough, so the following two ideas are our policy to improve ELR algorithm; (1) the termination condition of ELR is not “all of  $\lambda_{i,k}$  become zero” but “all  $\lambda_{i,k}$  become almost zero” and (2) make  $\tilde{c}_{k,k}$  small roughly but not severely. Based on these two ideas, our proposal is to use a floor function instead of a rounding function in Eq. 13. With this function,  $\lambda_{i,k}$  becomes almost zero faster than original. Note that the BER performance also decrease if the initial orthogonality is low. After all, we propose to use Eq. 16 instead of Eq. 13.

$$\lambda_{i,k} = \begin{cases} -\left\lfloor \frac{\tilde{c}_{i,k}}{\tilde{c}_{i,i}} \right\rfloor & (\tilde{c}_{k,k} > \alpha) \\ -\left\lfloor \frac{\tilde{c}_{i,k}}{\tilde{c}_{i,i}} \right\rfloor & (\tilde{c}_{k,k} \leq \alpha) \end{cases} \quad (16)$$

where  $\lfloor \cdot \rfloor$  denotes a floor function and  $\alpha$  denotes the threshold. In order to determine the threshold  $\alpha$ , investigate the PEP in Eq. 14 again. First, we assume that minimum orthogonality can be obtained if the PEP is smaller than  $10^{-4}$  by tradition. Since  $A(3.7) \cong 10^{-4}$ , the PEP is smaller than  $10^{-4}$  if the following condition is satisfied,

$$\tilde{c}_{k,k} < \frac{|e_{z_k}|^2}{27.38\sigma_w^2}. \quad (17)$$

The right term of Eq. 17 is threshold  $\alpha$ . After all, our proposal is shown in Table 2.

### 6. Experiments

We conducted some experiments with MATLAB 2018a to verify the superiority of our algorithm. Through the experiments, we assume that the modulation in transmitter is 64 QAM, the channel is additive white Gaussian noise and quasi-static Rayleigh fading, and only receiver has channel information.

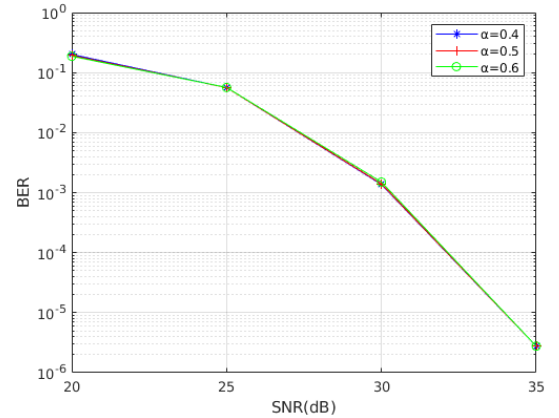
#### 6.1 Performance via SNR

First, the BER performance and the average number of

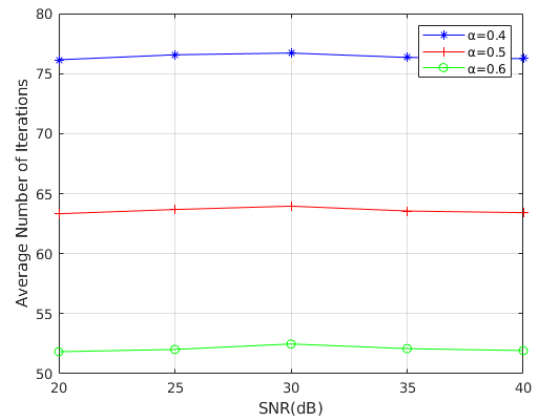
**Table 2** Our algorithm

Input: $\mathbf{H}$	Output: $\tilde{\mathbf{H}}, \mathbf{T}$
(1)	$\tilde{\mathbf{C}} = (\mathbf{H}^H \mathbf{H})^{-1}, \mathbf{T}' = \mathbf{I}_N$
(2)	<b>Do</b>
(3)	<b>Find the largest reducible</b> $\tilde{c}_{k,k}$
(4)	<b>If</b> $\tilde{c}_{k,k} > \alpha$ <b>then</b>
(5)	$\lambda_{i,k} = -\lfloor \tilde{c}_{i,k} / \tilde{c}_{i,i} \rfloor$
(6)	<b>Else</b>
(7)	$\lambda_{i,k} = -\lfloor \tilde{c}_{i,k} / \tilde{c}_{i,i} \rfloor$
(8)	<b>End if</b>
(9)	<b>If all</b> $\lambda_{i,k} = 0, \forall i \neq k$ <b>then, goto 11</b>
(10)	<b>Choose</b> $i = \arg \max_{i=1, i \neq k}^{N_T} \Delta_{i,k}$
(11)	$\mathbf{t}'_k = \mathbf{t}'_k + \lambda_{i,k} \mathbf{t}'_i$
(12)	$\tilde{\mathbf{c}}_k = \tilde{\mathbf{c}}_k + \lambda_{i,k} \tilde{\mathbf{c}}_i$
(13)	$\tilde{\mathbf{c}}^{(k)} = \tilde{\mathbf{c}}^{(k)} + \lambda_{i,k}^* \tilde{\mathbf{c}}^{(i)}$
(14)	<b>While(true)</b>
(15)	$\mathbf{T} = (\mathbf{T}'^{-1})^H, \tilde{\mathbf{H}} = \mathbf{H}\mathbf{T}$

iterations in a process of detection in our proposal with different threshold  $\alpha$  are shown in from Fig. 2 to Fig. 5.



**Fig. 2** BER via SNR for our algorithm with  $\alpha=0.4, 0.5,$  and  $0.6$ .



**Fig. 3** The Number for iterations via SNR for our algorithm with  $\alpha=0.4, 0.5,$  and  $0.6$ .

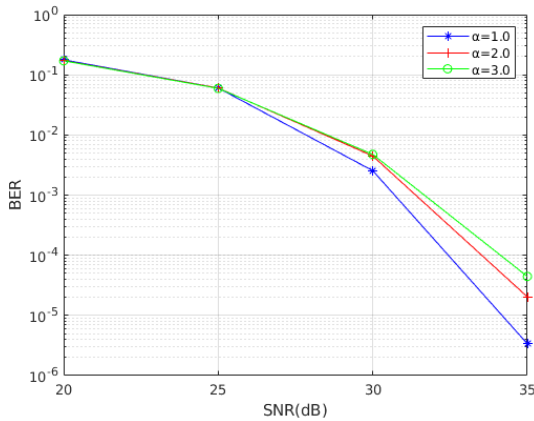


Fig. 4 BER via SNR for our algorithm with  $\alpha=1.0, 2.0$ , and  $3.0$ .

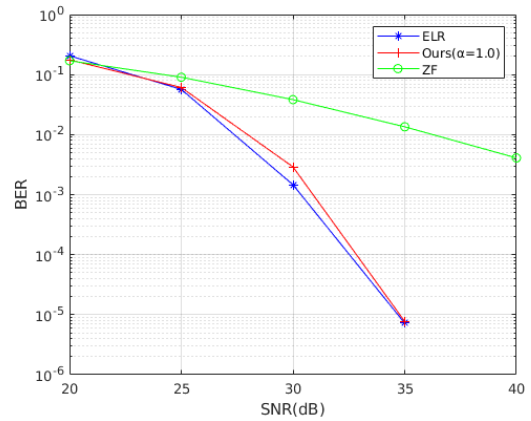


Fig. 6 BER via SNR for ZF and LR-aided ZF.

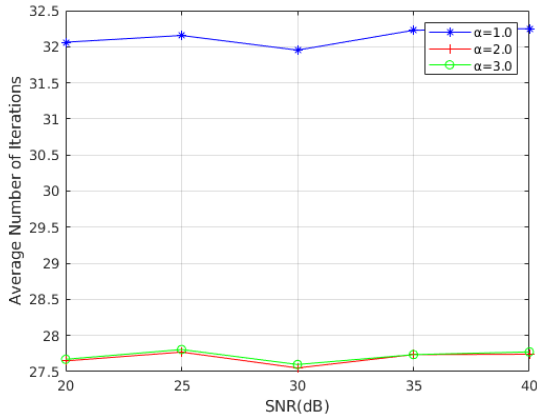


Fig. 5 The Number of iterations via SNR for our algorithm with  $\alpha=1.0, 2.0$ , and  $3.0$ .

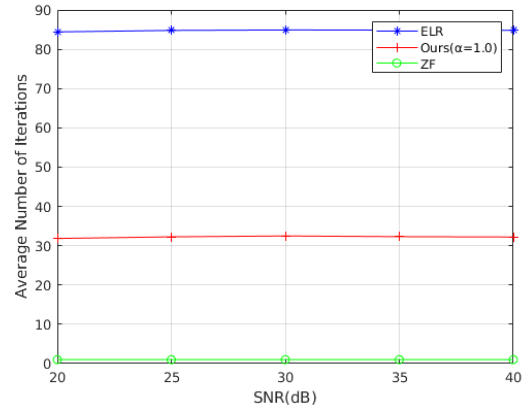


Fig. 7 The number of iterations via SNR for ZF and LR-aided ZF.

Fig. 2 and Fig. 3 are the case of  $\alpha = 0.4, 0.5$ , and  $0.6$ , and Fig. 4 and Fig. 5 are the case of  $\alpha = 1.0, 2.0$ , and  $3.0$ . In Fig. 2 and Fig. 4, the horizontal axis and vertical axis indicate the SNR and BER, respectively. In Fig. 3 and Fig. 5, the horizontal and vertical axis are the SNR and the average number of iterations in a process of detection, respectively. In these simulation, the number of transmit and receive antennas are 40, and SNR is 30dB. As you see, the number of iterations become smaller with almost the same BER performances when the threshold becomes smaller. Of course, the BER performance with  $\alpha=0.4$  is superior to that with  $\alpha=0.6$  or  $\alpha=0.5$ , but the difference between them is inconsiderable. In Fig. 4 and Fig. 5, you can also see that the number of iterations are not so different whether  $\alpha$  is 2.0 or 3.0. Comparing from Fig. 2 to Fig. 5 and other similar experiments,  $\alpha = 1.0$  looks good in view of tradeoff between the number of iteration and BER.

Next, the BER performance and the number of iteration of ZF, original ELR-aided ZF, and ours are shown in Fig. 6 and Fig. 7 whose horizontal axis and vertical axis indicate the SNR and BER, respectively. Both of the number of transmit and receive antennas are 40 and we use  $\alpha=1.0$  for ours experimentally. You can see that ELR-aided and ours-aided ZF defeat the original ZF, and that our algorithm shows almost the same BER performance as original ELR despite of

the smaller number of iterations of ours. Of course, the number of iteration of ZF is just 1 because it does not need any iteration.

### 6.2 Performance via the Number of Antennas

As the third experiment, we compared the performance of ZF, original-ELR-aided ZF, and ours with different number of antennas at SNR=30db. The result is shown in Fig. 8 and Fig. 9. The horizontal axes of both are the number of antennas, which is the same in a transmitter and a receiver. The vertical axis indicates BER in Fig. 8 and the number of iterations in Fig. 9. With 100 antennas for each of transmitter and receiver, the number of iterations of ours is 63.3% less than that of original even though the BER of ours is about 35.4% inferior to the other.

## 7. Conclusion

ZF method is a linear detection method but the BER is not so good. In order to improve the accuracy of linear detection, LR-aided linear detection methods is good solution. ELR-aided linear detection is one of them, and it is suit for massive MIMO systems because of its large number of antennas. However, ELR algorithm make the channel matrix more orthogonal than necessary, so we proposed

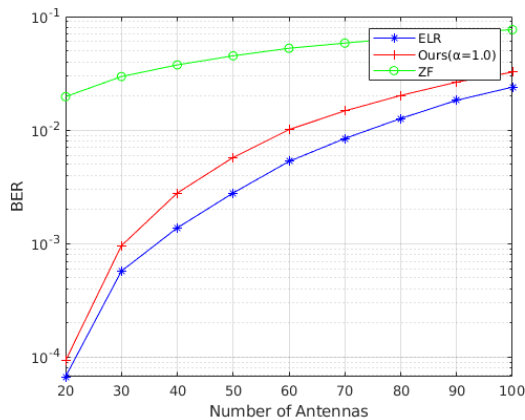


Fig. 8 BER via the number of antennas for ZF and LR-aided ZF.

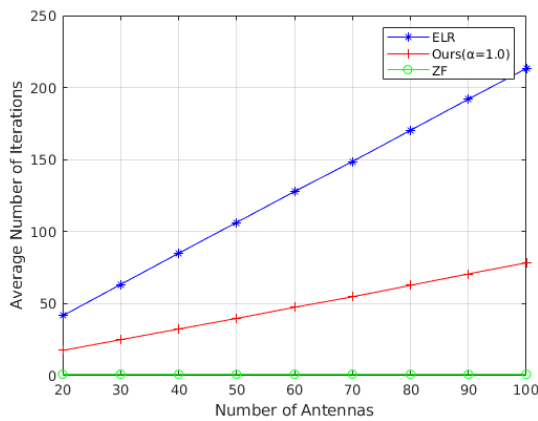


Fig. 9 The number of iteration via the number of antennas for ZF and LR-aided ZF.

improved ELR. The improvement point is just a change of the function in the algorithm.

We conducted some experiments in section 7, and we showed that our algorithm could reduce the average number of iterations in the process to detect a set of signals about a half of original although the BER of ours is not so inferior to that of original. It means that ours algorithm is more efficiency than original ELR. However, the threshold was determined experimentally in this paper, so we should determine the threshold  $\alpha$  logically.

**References**

[1] E. G. Larsson, "MIMO detection methods: How they work," IEEE Signal Processing Magazine, vol.26 no.3, pp.91-95, May 2009.  
 [2] J. Jalden and B. Ottersten, "On the complexity of sphere decoding in digital communications," IEEE Transaction on Signal Processing, vol.53, no.4, pp.1474-1484, April 2008.  
 [3] J. Jalden, D. Seethaler, and G. Matz, "Worst-and average-case complexity of LLL lattice reduction in MIMO wireless systems," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp.2685-2688, March 2008.  
 [4] Q. Zhou and X. Ma, "Designing low-complexity detectors for generalized SC-FDMA systems," 45th Annual Conference on

Information Science and Systems, pp.1-6, March 2011.  
 [5] Q. Zhou and X. Ma, "Element-Based Lattice Reduction Algorithms for Large MIMO Detection," IEEE Journal on Selected Areas in Communications, vol.31, no.2, February 2013.  
 [6] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," IEEE Transaction on Information Theory, vol.48, no.8, pp.2201-2214, August 2002.

(2018 年 11 月 5 日 受理)