# A High-Parallelism Detection Algorithm for Massive MIMO Systems

Huan Li, Xuying Zhao
*Institute of Automation, Chinese Academy of Sciences*
*University of Chinese Academy of Sciences*
*Beijing 100190, P. R. China*
{lihuan2013, zhaoxuying2012} @ia.ac.cn

Chen Guo, Donglin Wang
*Institute of Automation, Chinese Academy of Sciences*
*Chinese Academy of Sciences*
*Beijing 100190, P. R. China*
{chen.guo, donglin.wang} @ia.ac.cn

*Abstract*—**Due to the asymptotically orthogonal channel, minimum mean square error detection algorithm is near-optimal for uplink massive MIMO systems, but it involves matrix inversion with high complexity. This paper proposes a high-parallelism detection algorithm in an iterative way to avoid the complicated matrix inversion. The parallelism level is analyzed and convergence is proved in detail. The proposed algorithm can be implemented in a high level, which is equal to the max number of received data streams. The complexity can be reduced by one order of magnitude comparing with MMSE algorithm. Simulation results show that the proposed algorithm can closely match the performance of the MMSE algorithm with few number of iterations. It also outperforms Neumann Series approximation algorithm in terms of block error rate (BLER) performance with same number of iterations.**

***Keywords: Massive MIMO; Signal Detection; High Parallelism; Matrix Blocking***

## I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) is considered as one of the key technologies of 5G system, since it can make use of the large-scale antenna array to enhance the spectrum efficiency and energy efficiency [1][2]. Nevertheless, massive MIMO confronts some stressful problems for baseband signal processing, due to the expanded number of antennas [3]. An urgent problem is signal detection.

Linear detection algorithm enjoys better performance and lower complexity comparing with the traditional maximum likelihood (ML) algorithm based on the posteriori probability [4]. Therefore, linear detection algorithm, such as zero forcing (ZF) algorithm or minimum mean square error (MMSE) algorithm, is usually adopted in uplink system. Cholesky decomposition is required to acquire the matrix inversion for the above algorithms [4], and its computational complexity is $O\left(K^3\right)$, where K denotes the number of data streams. Due to the huge number of data streams supported by massive MIMO systems, linear detection algorithm is not suitable anymore. Therefore, there have been many attempts in simplifying the large-scale matrix inversion in recent years. Neumann series approximation (NSA) algorithm has become the most popular type of simplified algorithms [5], which transforms the matrix inversion into a series of matrix-vector multiplications. High-parallelism can be achieved by NSA algorithm, since the series expansions is a diagonal matrix, but the required expansion number is much large to match the performance of MMSE algorithm. The actual complexity has not been reduced. [6] proposed low-complexity algorithm based on Lanczos. The complexity is reduced comparing with the NSA algorithm, but the parallelism is also reduced. [7] proposed SOR method based on iteration, which enjoys better performance with a small number of iterations. However, in order to reduce the complexity, SOR algorithm can only be implemented serially, which is not suitable for popular single instruction multiple data (SIMD) processor. Furthermore, none of the above algorithms are optimized for divisions, but the delay in hardware caused by divisions is not negligible.

In this paper, a high-parallelism near-optimal algorithm based on matrix blocking is proposed. First of all, we raise the constitute method of the iterative matrix. Secondly, we prove the convergence of the proposed algorithm due to the features of the channel matrix of massive MIMO systems. Then the parallel computation method is described, as well as the complexity is analyzed. The parallelism can be the same as the matrix dimension, which is much higher than the above algorithms except NSA algorithm. Simultaneously, the complexity can be reduced by one order of magnitude comparing with MMSE algorithm. Simulation results show that the proposed algorithm can approximately match the MMSE algorithm with a small number of iterations, which is much less than the NSA algorithm with same level of parallelism.

The rest of the paper is constructed as follows. The massive MIMO system model is briefly presented in Section II. Section III proposes the parallelism algorithm. The convergence proof and parallel implementation method are also provided. In Section IV, simulation results and analysis are given. Finally, conclusions are drawn in Section V.

## II. SYSTEM MODEL

A multi-user massive MIMO system is presented in Fig. 1. Assuming that each user is a single antenna transmission, the base station contains N antennas and receives K data streams from K users. In massive MIMO system, $N \gg$ is usually adopted. Let $\boldsymbol{X}$ denote the transmitted signals, and the received signals from all the antennas can be denoted as $\boldsymbol{Y}$.

978-1-5090-3025-5/17/$31.00 ©2017 IEEE

Let $\boldsymbol{H} \in \mathbb{C}$ denote the channel impulse response matrix, and $\boldsymbol{N}$ denotes complex Gaussian noise, whose entries are independently and identically distributed (i.i.d) and follow the distribution $\boldsymbol{N}\left(0, \delta^2\right)$. A system model of multi-user massive MIMO can be considered as (1).

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{N} \tag{1}$$

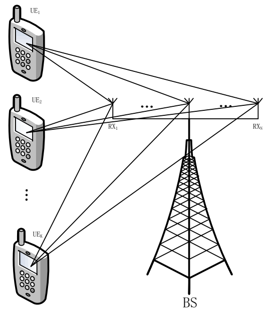


Fig. 1. Multi-user massive MIMO in the uplink.

In the light of MMSE algorithm [3], the estimate result of transmitted signal $\hat{\boldsymbol{X}}$ can be indicated as (2)

$$\hat{\boldsymbol{X}} = \left(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \delta^2\boldsymbol{I}_x\right)^{-1}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{Y} \tag{2}$$

(2) can be transformed into (3), where $\boldsymbol{b} = \boldsymbol{H}^{\mathrm{H}}\boldsymbol{Y}$ and $\boldsymbol{A} = \boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \sigma^2\boldsymbol{I}_x$ represents the matched-filter and filter.

$$\hat{\boldsymbol{X}} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{3}$$

The inversion of filter matrix $\boldsymbol{A}$ is calculated firstly and then multiplies the matched filter to estimate the result in MMSE algorithm. The calculation of the matrix $\boldsymbol{A}^{-1}$ is based on Cholesky decomposition and the complexity contained is $O\left(K^3\right)$, where K is usually 8, 16 or higher for massive MIMO systems to obtain higher system capacity. Based on that, the complexity of MMSE algorithm is too high to realize for hardware. Meanwhile, the Cholesky decomposition cannot be implemented in parallel. It is necessary to obtain the estimate result in other ways.

## III. A HIGH-PARALLELISM ALGORITHM

We transform (3) into the following system of linear equations (4) to prevent the complex matrix inversion. $\hat{\boldsymbol{X}}$ can be obtained in iterative way by dividing $\boldsymbol{A}$ into a sum of two matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$, where $\boldsymbol{P}$ is a nonsingular matrix.

$$\boldsymbol{A}\hat{\boldsymbol{X}} = \boldsymbol{b} \tag{4}$$

$$\boldsymbol{A} = \boldsymbol{P} + \boldsymbol{Q} \tag{5}$$

Due to (4)(5), the equation (6) can be acquired

$$\boldsymbol{X} = -\boldsymbol{P}^{-1}\boldsymbol{Q}\boldsymbol{X} + \boldsymbol{P}^{-1}\boldsymbol{b} \tag{6}$$

The solution of the system of linear equations in iterative way can be obtained as the following form

$$\begin{aligned} \boldsymbol{X}_{k+1} &= -\boldsymbol{P}^{-1}\boldsymbol{Q}\boldsymbol{X}_k + \boldsymbol{P}^{-1}\boldsymbol{b} \\ s.t.\ \lim_{x\to\infty}\left(-\boldsymbol{P}^{-1}\boldsymbol{Q}\right)^k &= 0 \end{aligned} \tag{7}$$

According to (7), matrix $\boldsymbol{P}$ is required to meet the convergence condition, and the inversion of $\boldsymbol{P}$ can be simply acquired. In addition, the iteration number required is small in order to reduce the complexity.

In this paper, we construct the iterative matrix $\boldsymbol{P}$ by splicing several block matrices, which are obtained by partitioning the matrix $\boldsymbol{A}$. The proposed algorithm can be implemented in hardware with high level of parallelism, and the computational complexity can be reduced to $O\left(K^2\right)$. In addition, by utilizing the properties of the $2\times 2$ matrix inversion, the number of division can be reduced by half compared with the algorithms mentioned in Section Ⅰ.

Matrix $\boldsymbol{P}$ is constructed as following steps. Firstly, we divide $\boldsymbol{A}$ into several $2\times 2$ sub-matrices. The number of sub-matrices in each line after partitioning is $M{=}K\ /\ 2$. Based on that, the original matrix $\boldsymbol{A}$ is divided into $M \times M$ sub-matrices. $\boldsymbol{P}$ is the result of combining all the M diagonal sub-matrices. The construction results of matrix $\boldsymbol{P}$ can be obtained as (8). Where $p_{i,j}$ and $a_{i,j}$ denote the values of the i-th row and j-th column elements of $\boldsymbol{P}$ and $\boldsymbol{A}$, respectively. Figure 2 shows the $\boldsymbol{P}$ when the matrix dimension is 8.

$$p_{i,j} = \begin{cases} a_{i,j},\ j-1 \le i \le j, & i \bmod 2 = 1 \\ a_{i,j},\ j \le i \le j+1, & i \bmod 2 = 0 \\ 0,\text{others} & \end{cases} \tag{8}$$

$$\begin{bmatrix} a_{1,1} & a_{1,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ a_{2,1} & a_{2,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{3,3} & a_{3,4} & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{4,3} & a_{4,4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{5,5} & a_{5,6} & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{6,5} & a_{6,6} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_{7,7} & a_{7,8} \\ 0 & 0 & 0 & 0 & 0 & 0 & a_{8,7} & a_{8,8} \end{bmatrix}$$

Fig. 2. The iterative matrix $\boldsymbol{P}$ (K=8, M=4, m=2).

### A. Convergence proof

First of all, we consider $\lambda_n$ as an arbitrary eigenvalue of matrix $-\boldsymbol{P}^{-1}\boldsymbol{Q}$ and $\boldsymbol{v}^{\mathrm{T}}$ as the corresponding eigenvector. (9) can be obtained according to their definitions. It can be represented as (10) based on (5)

$$-\boldsymbol{P}^{-1}\boldsymbol{Q}\boldsymbol{v}^{\mathrm{T}} = \lambda_n\boldsymbol{v}^{\mathrm{T}} \tag{9}$$

$$\boldsymbol{v}\left(2\boldsymbol{P} - 2\boldsymbol{A}\right)\boldsymbol{v}^{\mathrm{T}} = 2\lambda_n\boldsymbol{v}\boldsymbol{P}\boldsymbol{v}^{\mathrm{T}} \tag{10}$$

We denote $\boldsymbol{C} = \boldsymbol{P} - \boldsymbol{Q}$, (11) can be obtained based on (5)

$$2\boldsymbol{P} - 2\boldsymbol{A} = \boldsymbol{C} - \boldsymbol{A}$$
$$2\boldsymbol{P} = \boldsymbol{C} + \boldsymbol{A} \tag{11}$$

According to (10) (11), (12) can be acquired. Furthermore, (12) can be expressed as (13)

$$\boldsymbol{v}\left(\boldsymbol{C} - \boldsymbol{A}\right)\boldsymbol{v}^{\mathrm{T}} = \lambda_n \boldsymbol{v}\left(\boldsymbol{C} + \boldsymbol{A}\right)\boldsymbol{v}^{\mathrm{T}} \tag{12}$$

$$\left(1 - \lambda_n\right)\boldsymbol{v}\boldsymbol{C}\boldsymbol{v}^{\mathrm{T}} = \left(1 + \lambda_n\right)\boldsymbol{v}\boldsymbol{A}\boldsymbol{v}^{\mathrm{T}} \tag{13}$$

Massive MIMO systems contains special features that the column vectors of the channel matrix are asymptotically orthogonal, which is different from the LTE-A system. Considering $\boldsymbol{A} = \boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + \sigma^2 \boldsymbol{I}_x$, since $\sigma^2$ is greater than or equal to 0, at the same time, $\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}$ is a Hermite matrix, and diagonal elements of $\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}$ are positive, then $\boldsymbol{A}$ is a strictly diagonally dominant matrix as well as a positive definite matrix. The expression (14) is satisfied according to the definition of such a strictly diagonally dominant matrix. Meanwhile, the expression (15) is satisfied according to the property of positive definite matrix.

$$\left\| a_{j,j} \right\| > \sum_{i=1, i\neq j}^{K} \left\| a_{j,i} \right\| \tag{14}$$

$$\boldsymbol{v}\boldsymbol{A}\boldsymbol{v}^{\mathrm{T}} > 0 \tag{15}$$

Due to $\boldsymbol{C} = \boldsymbol{P} - \boldsymbol{Q}$, all the diagonal elements of $\boldsymbol{A}$ are kept, then (14) is satisfied. Thus $\boldsymbol{C}$ is a strictly diagonally dominant matrix. Meanwhile, as the diagonal elements of $\boldsymbol{C}$ are positive, $\boldsymbol{C}$ is a Hermite positive definite matrix, which makes (16) be satisfied

$$\boldsymbol{v}\boldsymbol{C}\boldsymbol{v}^{\mathrm{T}} > 0 \tag{16}$$

According to the expression (13)(15)(16), (17) can be acquired. $\left|\lambda_n\right| < 1$ can be obtained, which equals to the convergence condition. It can be concluded that the iteration process is converged using the proposed iterative matrix.

$$\left(1 - \lambda_n\right)\left(1 + \lambda_n\right) > 0 \tag{17}$$

*B. Parallel Implementation Method*

The estimate result of transmitted signal can be obtained in (18).

$$\boldsymbol{X}_{k+1} = \boldsymbol{P}^{-1}\left(\boldsymbol{b} - \boldsymbol{Q}\boldsymbol{X}_k\right) \tag{18}$$

First of all, it can be found from (18) that in the k-th iteration, one multiplication of the $K \times K$ matrix $\boldsymbol{Q}$ and the $K \times 1$ vector $\boldsymbol{X}_k$ is involved in the proposed algorithm. Since there are equal amounts of $\left(K - 2\right)$ nonzero elements in each row in $\boldsymbol{Q}$, the matrix-vector multiplication can be implemented in parallel by K multiplication of vector-vector, which make the operation can be obtained a high level of parallelism. The complexity of one multiplication of matrix-vector is $\left(K - 2\right) \times K$.

Secondly, it is necessary to consider the inversion of $\boldsymbol{P}$. $\boldsymbol{P}$ is constructed by the partitioning method. Based on that, the inversion can be implemented in parallel by M inversion of $2 \times 2$ matrix. The inversion of $2 \times 2$ matrix can be obtained as the following form.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{19}$$

All the $2 \times 2$ matrix in this paper is a Hermite matrix, then the required number of multiplications is 4. One division is also required. Therefore, the multiplication and division times brought by all $2 \times 2$ matrix inversions are $2K$ and $K/2$.

Then, we consider the multiplication of $K \times K$ matrix $\boldsymbol{P}^{-1}$ and $K \times 1$ vector $\left(\boldsymbol{b} - \boldsymbol{Q}\boldsymbol{X}_k\right)$, which are obtained in the previous step. Similar to $\boldsymbol{Q}$, there are equal amounts of 2 nonzero elements in each row in $\boldsymbol{P}^{-1}$. Therefore, the multiplication of matrix-vector also can be implemented in parallel by K. The complexity of one multiplication of matrix-vector is $2K$.

It can be concluded that the parallelism of implementation in one iteration is K and the multiplications required is $K^2$. To obtain the inversion of $\boldsymbol{P}$, $K/2$ times of divisions and $2K$ times of multiplications are required.

TABLE I. PARALLELISM AND COMPLEXITY COMPARASION

| | | Proposed | NSA | SOR |
|---|---|---|---|---|
| mul | Iter-=2 | $K^2$+5K | $2K^{2-}$K | $2K^2$+2K |
| | Iter-=3 | $2K^2$+5K | $K^3$ | $3K^2$+2K |
| | Iter-=4 | $3K^2$+5K | $2K^3$-$K^2$ | $4K^2$+2K |
| | Iter-=5 | $4K^2$+5K | $3K^3$-$2K^2$ | $5K^2$+2K |
| Div | | K/2 | K | K |
| Par | | K | K | 1 |

Table 1 compares the parallelism and complexity of the proposed algorithm, NSA and SOR algorithm. Note that the complexity of MMSE algorithm is $O\left(K^3\right)$. We can conclude that the parallelism of the proposed algorithm and NSA is higher than the SOR algorithm, since SOR algorithm reduces the complexity by using the forward substitution method. The complexity of the proposed algorithm and SOR algorithm is lower than the NSA algorithm, which can be reduced from $O\left(K^3\right)$ to $O\left(K^2\right)$ whenever the number of iterations is. Additionally, proposed algorithm requires K/2 times of divisions, which is half of NSA algorithm and SOR algorithm.

In a word, the proposed algorithm is more suitable for hardware implementation for its high parallelism and low complexity.

## IV. SIMULATION RESULTS

The simulation performance of the massive MIMO system with different algorithms and different number of iterations is provided in this section. Turbo encoding and 16QAM are adopted for the coding and modulation scheme. The number of antennas at base station is 64, and the number of users is 8 in the simulation.

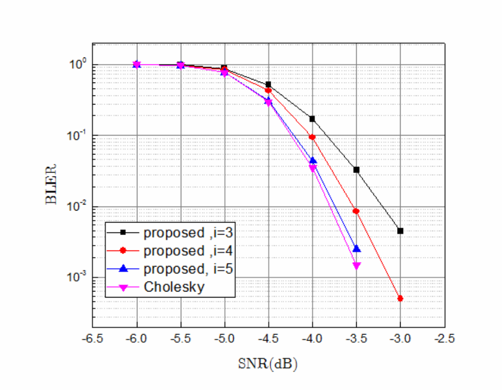


Fig.3. The simulation results of proposed algorithm

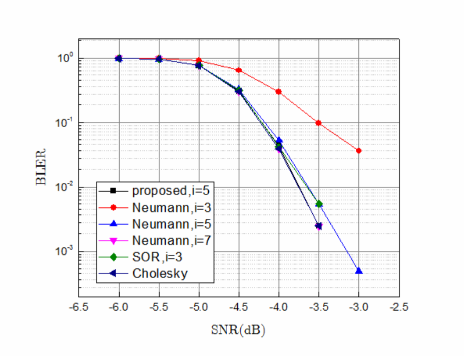


Fig.4. the performance comparison between the proposed algorithm, NSA algorithm and SOR algorithm

Figure 3 compares the performance of proposed algorithm with different numbers of iterations with MMSE algorithm as a standard. It can be obtained that the performance of the proposed algorithm is improved with enhanced number of iterations. When the number of iteration is 3, the performance gap between the proposed algorithm and the MMSE algorithm is 2.0dB. The performance of the proposed method is almost same as the standard result when the iteration number is 5.

Figure 4 compares the performance of the proposed algorithm, NSA algorithm and SOR algorithm. All the three algorithms can approach the standard results after several iterations. 7 times of iterations is required for NSA algorithm and only 3 times of iterations is required for the SOR algorithm. The proposed algorithm and SOR algorithm converge much faster than NSA algorithm. Taking into account the complexity, the multiplications required by proposed algorithm is $K^2$ times more than SOR algorithm, but K/2 times of divisions is reduced. The delay in the hardware caused by the above two factors is approximately equal. Although the complexity of the proposed algorithm and the SOR algorithm is almost the same, there is a big difference for the implementation when taking into account the impact of parallel implementation. Since the parallelism of the proposed algorithm is K times of the SOR algorithm, the execution cycle can be reduced to 1 / K compared with the SOR algorithm in hardware.

## V. CONCLUSIONS

In this paper, a high-parallelism detection algorithm is proposed, which transforms the complex matrix inversion into a linear system of equation. The iteration matrix is constructed and the convergence is proved in detail. The parallel implementation method and complexity analysis shows that parallelism can be the same as the matrix dimension, which is higher than the SOR and Lanczos algorithms. The complexity can be reduced by one order of magnitude comparing with the traditional MMSE algorithm, while half of the divisions can be reduced comparing with existing simplified algorithms. Simulation results shows that the proposed algorithm can closely match the performance of MMSE algorithm with only a small number of iterations. The required number of iterations is near half comparing with the NSA algorithm. The equivalent hardware complexity is near as same as SOR algorithm, but the proposed algorithm can be implemented with high level of parallelism.

## REFERENCES

[1] H. NGO, E. LARSSON and T. MARZETTA, “Energy and spectral efficiency of very large multiuser MIMO systems,”. IEEE Communications Magazine, vol. 61,2012, pp.1436–1449, doi:10.1109/TCOMM.2013.020413.110848.

[2] R. W. HEALTH, A. LOZANO, T. L. MARZETTA, et al, “Five disruptive technology directions for 5G,” IEEE Communications Magazine,vol.52,2014, pp.74–80, doi:10.1109/MCOM.2014.6736746.

[3] F. RUSEK, D. PERSSON, B. K. LAU, et al, “Scaling up MIMO: Opportunities and challenges with very large arrays,” IEEE Signal Processing Magazine, vol. 30 2013, pp. 40–60, doi:10.1109/MSP.2011.2178495.

[4] Licai FANG, Lu XU and Defeng David HUANG, “Low Complexity Iterative MMSE-PIC Detection for Medium-Size Massive MIMO,” IEEE Wireless Communications Letters, vol.5, 2016, pp.108-111, doi:10.1109/LWC.2015.2504366.

[5] Chuan TANG, Cang LIU, Luechao YUAN, et al, “High Precision, Low Complexity Matrix Inversion Based on Newton Iteration for Data Detection in the Massive MIMO,” IEEE COMMUNICATIONS LETTERS, vol.20, 2016 pp. 490–493,doi:10.1109/LCOMM.2015.2514281.

[6] Huaibo Zhang, Guiqiang Peng and Leibo Liu , “Low Complexity Signal Detector Based on Lanczos Method for Large-Scale MIMO Systems,”. IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC), 2016, pp.6-9, doi: 10.1109/ICEIEC.2016.7589675

[7] Peng ZHANG, Leibo LIU, Guiqiang PENG, et al, “Large-scale MIMO detection design and FPGA implementations using SOR method,” IEEE International Conference on Communication Software and Networks (ICCSN), 2016, pp.206–210, doi:10.1109/ICCSN.2016.758665.