A comparative study of low-complexity MMSE signal detection for massive MIMO systems

Shufeng Zhao[1], Bin Shen[1], and Quan Hua[1]
[1] School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications
[e-mail: zhaoshuf@foxmail.com, shenbin@cqupt.edu.cn]

Received May 23, 2017; revised October 15, 2017; revised November 12, 2017; accepted December 4, 2017; published April 30, 2018

Abstract

For uplink multi-user massive MIMO systems, conventional minimum mean square error (MMSE) linear detection method achieves near-optimal performance when the number of antennas at base station is much larger than that of the single-antenna users. However, MMSE detection involves complicated matrix inversion, thus making it cumbersome to be implemented cost-effectively and rapidly. In this paper, we first summarize in detail the state-of-the-art simplified MMSE detection algorithms that circumvent the complicated matrix inversion and hence reduce the computation complexity from $\mathcal{O}(K^3)$ to $\mathcal{O}(K^2)$ or $\mathcal{O}(NK)$ with some certain performance sacrifice. Meanwhile, we divide the simplified algorithms into two categories, namely the matrix inversion approximation and the classical iterative linear equation solving methods, and make comparisons between them in terms of detection performance and computation complexity. In order to further optimize the detection performance of the existing detection algorithms, we propose more proper solutions to set the initial values and relaxation parameters, and present a new way of reconstructing the exact effective noise variance to accelerate the convergence speed. Analysis and simulation results verify that with the help of proper initial values and parameters, the simplified matrix inversion based detection algorithms can achieve detection performance quite close to that of the ideal matrix inversion based MMSE algorithm with only a small number of series expansions or iterations.

Keywords: simplified matrix inversion, classical iterative method, minimum mean square error, signal detection, massive MIMO

This work is supported by the Innovation Project of the Common Key Technology of Chongqing Science and Technology Industry (Grant no. cstc2015zdcy-ztzx40008).

http://doi.org/10.3837/tiis.2018.04.007 **ISSN : 1976-7277**

1. Introduction

Multiple-input multiple-output (MIMO) technology has been widely studied during the last two decades and already utilized in many wireless standards, such as LTE/LTE-A cellular network and IEEE 802.11n wireless LAN system, since it can significantly enhance the capacity and reliability of wireless systems on the premise of demanding no additional bandwidth or transmit power[1]-[2]. For conventional MIMO systems, the base station (BS) is typically equipped with only a few antennas simultaneously serving multiple users within the same time-frequency resource, and the corresponding improvement in spectral efficiency is thus still relatively modest [3]. In order to exploit the resources more efficiently, ameliorate the transmission rate and relieve the system interference, an emerging technology referred to as massive MIMO, which entails ten to hundreds of antennas at BS, was proposed in recent years [4]-[5]. Extra antennas amounted at BS help focus energy into ever-smaller regions of space to bring huge improvements in throughput and radiated energy efficiency [6]. Massive MIMO scales up conventional MIMO by possibly orders of magnitude and has been regarded as an enabler for the development of future broadband wireless networks [7]-[8].

However, some challenging problems have to be solved to realize such attractive merits of massive MIMO in practice [9]. One of them is developing low-complexity signal detection algorithms in uplink when the number of single-antenna users is becoming extremely large [5]. A number of signal detection algorithms that work efficiently in conventional small-scale MIMO systems fail in massive MIMO systems on account of computational complexity or performance. For example, the complexity of maximum likelihood (ML) detector, which is optimal among the hard decision methods, grows exponentially with the modulation order and the number of transmit antennas. The tabu search (TS) [10] and the fixed-complexity sphere decoding (FSD) [11] algorithms were put forward to obtain the close optimal ML detection performance with reduced complexity, but their complexity becomes not affordable when the configuration of MIMO system is large or the modulation order is high [12]. Encountering the tremendously enlarged number of antennas, one has no choice but to turn to linear signal detection algorithms on the basis of zero-forcing (ZF) and minimum mean square error (MMSE) criteria, because of their relatively low complexity and satisfactory bit error rate (BER) performance for multiuser massive MIMO systems [5]. However, these methods inevitably involve complicated matrix inversion due to the large dimensions of massive MIMO systems, resulting in highly burdensome complexity in practice. Very recently, many efforts have been dedicated to relieving the complexity of matrix inversion for practical detector design.

Neumann series expansion was proposed in [13] to replace the matrix inversion in MMSE detection, the performance and computational complexity of which scaled with the number of selected terms of Neumann series. However, when the number of selected terms is larger than two, the complexity of Neumann series expansion method was almost the same as that of exact matrix inversion. With an aim to decrease the complexity when the number of selected terms is three, the modified Neumann series expansion was proposed in [14]. In [15], multistage linear receiver (MLR) was utilized to approximate the exact matrix inversion and it has been shown that by using the optimized coefficient at each term it performs well with low number of terms. Different from the Neumann series expansion, Newton iterative approximation was proposed to approach the matrix inversion by a certain number of iterations in [16].

Unlike the above-mentioned methods mainly relying on the philosophy of matrix inversion

approximation, Richardson iteration (RI) method was proposed in [17] to avoid complicated matrix inversion by directly solving the linear equation of signal detection in an iterative manner, but there is plenty of room to improve the tradeoff between the signal detection performance and computational complexity. In [18] and [19], successive over-relaxation (SOR) and symmetric successive over-relaxation (SSOR) methods were respectively proposed to refrain from matrix inversion and the SSOR is an improved version of SOR. In addition, the work in [20] put forward conjugate gradient (CG) method to shun the costly inversion. Approximate message passing (AMP) method was proposed in [21] for iterative massive MIMO detection and decoding, but it still counted on the iterative updating of mean and variance between symbol nodes and observation nodes, ending up with a relatively complex process.

In this paper, aiming to further relieve the complexity of the aforementioned existing algorithms, we first apply the MLR method as the first iteration of Newton iterative approximation to reduce the complexity, and subsequently propose a weighted symmetric SOR (WSSOR) method on the basis of SOR and SSOR method to estimate the transmitted vector. Moreover, considering the series of approximate message passing (AMP) algorithms were initially proposed for solving a least absolute shrinkage and selection operator (LASSO) problem in the domain of compression sensing [22]-[23], we present a new set of message passing rules for the AMP method to obtain a simplified AMP (SAMP) algorithm. The SAMP method avoids the iterative updating of the mean and variance parameters in detection, which has a substantial reduction in computational complexity and still maintains good performance. Based on the eigenvalue decomposition of the interference and noise covariance matrix, the conventional minimum mean square error-interference rejection combining (MMSE-IRC) signal detection algorithm was proposed in [24] to exploit a dimension-reduction technique to reduce the computation-intensive of the matrix inversion compared with the conventional algorithm.

In order to give a more comprehensive study of the state-of-the-art simplified algorithms for implementing the MMSE detector, we classify these simplified matrix inversion algorithms into two categories which are respectively based on approximate matrix inversion and the classical iterative linear equation solving method. A detailed comparative analysis are carried out among them. In the sequel, we also address a variety of optimization schemes for the existing algorithms, such as presenting a more proper solution of initial values and relaxation parameter, reconstructing the effective noise variance. Finally, the performance and computational complexity of each type of the detection algorithms are verified by computer simulations. Among the simplified algorithms based on matrix inversion approximation, the MLR method is optimal; as for the simplified algorithms based on linear equation solving, the SAMP method and the SSOR iterative algorithm are the most desirable algorithms and the SAMP method achieves the best trade-off between performance and computational complexity over all the simplified algorithms.

The rest of paper is organized as follows: Section II introduces the typical massive MIMO system model. Low-complexity signal detectors based on polynomial expansion and linear equation solving are presented in Section III and Section IV respectively. Section V specifies the optimization schemes for the simplified algorithms. The simulation results are presented in Section VI to illustrate the effectiveness of the proposed detectors. Section VII provides a summary of our findings and concludes the paper.

Notations: Lower-case and upper-case boldface letters are used to represent column vectors and matrices, respectively. The superscripts T, $*$, H and -1 stand for the transpose, conjugate, conjugate-transpose and inverse of matrix, respectively. Tr(.) denotes the matrix trace and $\boldsymbol{I}_K$ is a $K \times K$ identity matrix. The operator $\|\cdot\|$ denotes the matrix norm and (A,B)

represents an inner product operation. $\Re(\cdot)$ and $\Im(\cdot)$ are the real and imaginary parts of a complex number, respectively.

2. Massive MIMO System Model

We consider a well-known uplink multiuser massive MIMO system which is composed of N receive antennas at BS and K single-antenna user equipments. Usually we have the system configuration as $N \geq K$, e.g., $N = 128$ and $K = 16$. Let $\boldsymbol{s}_c = [s_1, s_2, \cdots, s_K]^T$ represent the transmit vector consisting of the modulation symbols transmitted simultaneously by all users in the same time-frequency resources, where $s_k \in \mathbb{B}$ is the symbol transmitted from the *k*-th user, $\mathbb{B}$ is the modulation alphabet, and s_k is assumed to be with zero mean and finite variance σ_s^2. Let $\boldsymbol{H}_c \in \mathbb{C}^{N \times K}$ represent the channel coefficient matrix, whose entries are assumed to be independently and identically distributed with zero mean and variance σ_h^2. Therefore, the received $N \times 1$ signal vector $\boldsymbol{y}_c$ at BS can be denoted as

$$\boldsymbol{y}_c = \boldsymbol{H}_c \boldsymbol{s}_c + \boldsymbol{n}_c, \tag{1}$$

where subscript *c* stands for the complex-valued, and $\boldsymbol{n}_c$ is an $N \times 1$ additive white Gaussian noise (AWGN) vector whose entries follow the complex Gaussian distribution $\mathcal{CN}(0, \sigma_z^2)$.

Focusing on the uplink signal detection, when the subscript is dropped for convenience, we may rewrite the complex-valued system model of equation (1) in the real domain as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} + \boldsymbol{n}, \tag{2}$$

where $\boldsymbol{H} \in \mathbb{R}^{2N \times 2K}$, $\boldsymbol{y} \in \mathbb{R}^{2N}$, and $\boldsymbol{n} \in \mathbb{R}^{2N}$, which are respectively given by

$$\boldsymbol{H} = \begin{bmatrix} \Re(\boldsymbol{H}_c) & -\Im(\boldsymbol{H}_c) \\ \Im(\boldsymbol{H}_c) & \Re(\boldsymbol{H}_c) \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} \Re(\boldsymbol{y}_c) \\ \Im(\boldsymbol{y}_c) \end{bmatrix}, \quad \boldsymbol{s} = \begin{bmatrix} \Re(\boldsymbol{s}_c) \\ \Im(\boldsymbol{s}_c) \end{bmatrix}, \quad \boldsymbol{n} = \begin{bmatrix} \Re(\boldsymbol{n}_c) \\ \Im(\boldsymbol{n}_c) \end{bmatrix}. \tag{3}$$

The task of massive MIMO signal detection at BS is to detect the transmitted signal vector $\boldsymbol{s}$ on the basis of the received signal vector $\boldsymbol{y}$, and the estimation of $\hat{\boldsymbol{s}}$ coming from K different users can be achieved in MMSE detection scheme as

$$\hat{\boldsymbol{s}} = (\boldsymbol{H}^H \boldsymbol{H} + \sigma^2 \boldsymbol{I}_{2\mathrm{K}})^{-1} \boldsymbol{H}^H \boldsymbol{y} = \boldsymbol{W}^{-1} \boldsymbol{y}_{\mathrm{MF}}, \tag{4}$$

where the estimated variance σ^2 is assumed as $\sigma_z^2 / (\sigma_h^2 \sigma_s^2)$, the matched-filter output $\boldsymbol{y}_{\mathrm{MF}}$ can be interpreted as $\boldsymbol{y}_{\mathrm{MF}} = \boldsymbol{H}^H \boldsymbol{y}$, and the MMSE filtering matrix $\boldsymbol{W}$ is described as

$$\boldsymbol{W} = \boldsymbol{G} + \sigma^2 \boldsymbol{I}_{2\mathrm{K}}, \tag{5}$$

where $\boldsymbol{G} = \boldsymbol{H}^H \boldsymbol{H}$ stands for the Gram matrix. It is worth noting that the exact computational complexity of matrix inversion $\boldsymbol{W}^{-1}$ requires $\mathcal{O}(K^3)$, which increases exponentially with the

number of single-antenna users.

Due to the fact that $\boldsymbol{W}$ approaches the diagonal matrix when the number of antennas at BS is much larger than the number of singer-antenna users $(N \gg K)$ in massive MIMO systems, it has been suggested that we can simply approximate the matrix inversion by the inversion of its diagonal elements, meaning that we approximate $\boldsymbol{W}$ as

$$\boldsymbol{W} \approx \boldsymbol{D}^H \boldsymbol{D} + \sigma^2 \boldsymbol{I}_{2\mathrm{K}}, \tag{6}$$

where $\boldsymbol{D}$ is the main diagonal of $\boldsymbol{W}$. However, this rough approximation method will cause a huge performance loss in actual system configuration.

3. Low-Complexity signal detector based on approximate matrix inversion

A key property of massive MIMO systems is the orthogonality among channels when the number of antennas at BS is much larger than that of the single-antenna users $(N \gg K)$. Instead of dealing with the exact matrix inversion operation, which has a prohibitively high complexity, and taking advantage of the channel orthogonal property in MMSE detection, we can simply approximate the matrix inverse by the inversion of its diagonal elements only, i.e. $\boldsymbol{W}^{-1} \approx \boldsymbol{D}^{-1}$. However, in actual system configuration, this rough approximation method causes a huge performance degradation. Hence, in this section, we will mainly introduce some simplified algorithms based on the philosophy of matrix inversion approximation to cut down the computational complexity effectively.

3.1 Neumann series expansion

According to the Neumann series expansion [13], the inversion of matrix $\boldsymbol{W}$ can be simplified as

$$\boldsymbol{W}^{-1} = \sum_{n=0}^{\infty} \left(\boldsymbol{X}^{-1} (\boldsymbol{X} - \boldsymbol{W}) \right)^n \boldsymbol{X}^{-1}, \tag{7}$$

where the series expansion is guaranteed to converge if $(\boldsymbol{I}_{2\mathrm{K}} - \boldsymbol{X}^{-1}\boldsymbol{W}) = \boldsymbol{0}_{2\mathrm{K}}$ is fulfilled. By decomposing $\boldsymbol{W}$ such that $\boldsymbol{W} = \boldsymbol{D} + \boldsymbol{E}$, where $\boldsymbol{D}$ and $\boldsymbol{E}$ are the main diagonal and hollow component of $\boldsymbol{W}$ respectively, we can compute the first t terms of the Neumann series expansion as follows:

$$\boldsymbol{W}^{-1} = \sum_{n=0}^{t-1} \left(-\boldsymbol{D}^{-1}\boldsymbol{E} \right)^n \boldsymbol{D}^{-1}, \tag{8}$$

where we replace $\boldsymbol{X}$ in (7) with $\boldsymbol{D}$. It is worth mentioning that if $\lim_{t\to\infty} (-\boldsymbol{D}^{-1}\boldsymbol{E})^t = \boldsymbol{0}_{2\mathrm{K}}$, the series expansion in (8) holds.

For small values of t, the Neumann series expansion can be computed at low computational complexity. For instance, we can obtain $\boldsymbol{W}_1^{-1} = \boldsymbol{D}^{-1}$ when $t = 1$, whose computational complexity only scales with $\mathcal{O}(K)$ operations; this is a significant complexity

reduction in contrast to the $\mathcal{O}(K^3)$ required by dealing with MMSE exact matrix inversion. When $t=2$, we obtain $\boldsymbol{W}_2^{-1}=\boldsymbol{D}^{-1}-\boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{D}^{-1}$, whose computational complexity is reduced to $\mathcal{O}(K^2)$. When $t=3$, we have $\boldsymbol{W}_3^{-1}=\boldsymbol{D}^{-1}-\boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{D}^{-1}-\boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{D}^{-1}$, whose complexity scales with $\mathcal{O}(K^3)$, which is the same as that of an exact matrix inversion in MMSE detector.

To alleviate the computation complexity of the Neumann series expansion with $t=3$, the work in [14] proposed a modified Neumann series expansion which leverages that $\boldsymbol{W}^{-1}$ has the property of being diagonally dominant. The magnitude of each non-diagonal element $|X_{i,j}|$ in matrix is smaller than that in $\boldsymbol{D}^{-1}$ and $\boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{D}^{-1}$. In a nutshell, the magnitude of each non-diagonal element $X_{i,j}$ where $|i-j|>0$ tends to be zero. On account of the above facts, the modified Neumann series expansion can be expressed as follows:

$$\tilde{\boldsymbol{W}}^{-1}=\boldsymbol{D}^{-1}-\boldsymbol{D}^{-1}\boldsymbol{E}\boldsymbol{D}^{-1}+\tilde{\boldsymbol{X}}, \tag{9}$$

$$\tilde{\boldsymbol{X}}=\mathrm{diag}(\boldsymbol{X})+\sum_{1}^{T}\mathrm{diag}_T(\boldsymbol{X}),\quad (T>|i-j|), \tag{10}$$

where $\tilde{\boldsymbol{X}}$ represents the approximated $\boldsymbol{X}$ matrix and T is the predefined offset between the row index (i) and the column index (j) of $X_{i,j}$. Since the modified approach uses only numerically dominant elements of channel matrix for a low complexity matrix inversion, the computational complexity of the modified Neumann series expansion is lowered to $\mathcal{O}(K^2)$.

3.2 Multistage linear receiver

Although Neumann series expansion [13] and modified Neumann series expansion [14] were proposed to reduce the computational complexity in some sense, their performance still has much room for improvement. Motivated by this fact, the multistage linear receiver (MLR) was introduced to approximate the inversion of complicated matrix [15] and the pivotal idea is to substitute the matrix inversion with a polynomial by which (6) can be represented by

$$\boldsymbol{W}^{-1}=(\boldsymbol{G}+\sigma^2\boldsymbol{I}_{2\mathrm{K}})^{-1}=\sum_{s=0}^{t}\omega_s(\boldsymbol{G})^s, \tag{11}$$

where ω_s is the optimal weight of term s and t is the number of terms we choose.

Let $\boldsymbol{\omega}=[\omega_0,\omega_1,\cdots,\omega_s,\cdots,\omega_t]^T$, and the optimal weight vector can be denoted as

$$\boldsymbol{\omega}=\boldsymbol{\Phi}^{-1}\boldsymbol{c}, \tag{12}$$

where $\boldsymbol{\Phi}$ is a $(t+1)\times(t+1)$ square matrix with elements

$$\Phi_{i,j}=\mathrm{Tr}\left[(\boldsymbol{G})^{i+j+2}\right]+\sigma^2\mathrm{Tr}\left[(\boldsymbol{G})^{i+j+1}\right], \tag{13}$$

and $\boldsymbol{c}$ is a $(t+1)\times 1$ column vector with elements

$$c_i = \mathrm{Tr}\left[(\boldsymbol{G})^{i+1}\right]. \tag{14}$$

For the MLR method represented by equation (11) and (12), the computation complexity of calculating the optimal weights is $(t+1)^2$, because it refers to a t+1 dimensional matrix inversion. In addition, the computation complexity of constructing the MLR method can be reduced to $4tK^2$ by employing recursive method and we define the recursive vector as

$$\boldsymbol{r}_s = \omega_s \left(\boldsymbol{G}\right)^s \boldsymbol{y}_{\mathrm{MF}}, \tag{15}$$

$$\boldsymbol{r}_{s+1} = \left[\frac{\omega_{s+1}}{\omega_s}\boldsymbol{G}\right]\boldsymbol{r}_s. \tag{16}$$

It is clear that each time we only require to calculate a vector-matrix multiplication. Thus, the total computation complexity of the MLR method is $(t+1)^2+4tK^2$, whose computational complexity is decreased to $\mathcal{O}(K^2)$.

3.3 Newton iterative approximation

As can be seen from the above description, the core ideas underlying Neumann series expansion and MLR are both mathematically based on polynomial expansion. It is coincidentally interesting that the Newton iterative approximation proposed in [16] also utilized iterative method to approximate $\boldsymbol{W}^{-1}$. The proposed algorithm originates from the Taylor series which only takes the 1-order into consideration, and therefore improve the precision by increasing the iterative order.

If $\boldsymbol{W}_0^{-1}$ is the first time estimation of $\boldsymbol{W}^{-1}$, then the t-th iterative estimation can be represented by

$$\boldsymbol{W}_t^{-1} = \boldsymbol{W}_{t-1}^{-1}\left(2\boldsymbol{I}_{2\mathrm{K}} - \boldsymbol{W}\boldsymbol{W}_{t-1}^{-1}\right). \tag{17}$$

In order to ensure convergence, $\boldsymbol{W}_0^{-1}$ must meet the condition

$$\| \boldsymbol{I}_{2\mathrm{K}} - \boldsymbol{W}\boldsymbol{W}_0^{-1}\| < 1, \tag{18}$$

then $\boldsymbol{W}_t^{-1}$ in equation (17) quadratically converges to $\boldsymbol{W}^{-1}$.

From equation (17), we see that the computational complexity of Newton method is extremely high, which is $8K^3 + 4K$ for each iteration, because the initial value of Newton method is related to the matrix inversion $\boldsymbol{W}_0^{-1}$. Therefore, an appropriate selection of initial input $\boldsymbol{W}_0^{-1}$ will obviously affect the approximation accuracy. Inspired by Neumann series expansion, we can decompose $\boldsymbol{W} = \boldsymbol{D} + \boldsymbol{E}$ and use $\boldsymbol{D}^{-1}$ as the initial value. We will find that the first two iterations are duplicate as Neumann series expansion.

The MLR method can be employed to calculate the first term for the initial value of Newton method as well, and we select the first term of MLR as the initial value for Newton method,

$$\boldsymbol{W}_0^{-1} = \omega_0 + \omega_1(\boldsymbol{H}^H \boldsymbol{H}), \tag{19}$$

where the computational complexity is also very high compared with the initial value $\boldsymbol{D}^{-1}$. However, with this initial value in equation (19), we can achieve the BER close to that of exact MMSE detection performance by just two iterations.

By nature of its iterative calculation mechanism, the convergence rate of Newton iterative approximation is fast, and meanwhile the precision and complexity can be merely dominated by the number of iterations.

4. Low-Complexity signal detector based on classical iterative equation solving method

For uplink massive MIMO systems, the MMSE filtering matrix $\boldsymbol{W}$ is symmetric positive definite when the number of antennas at BS is much larger than that of the single-antenna users ($N \gg K$), as proved in [17]-[18]. Inspired by this special property, equation (4) can be rewritten as $\boldsymbol{W}\hat{\boldsymbol{s}} = \boldsymbol{y}_{\mathrm{MF}}$, which converts the simplified matrix inversion problem into the one of linear equation solving

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{20}$$

where $\boldsymbol{A}$ denotes the symmetric positive definite matrix, $\boldsymbol{x}$ is the $N \times 1$ solution vector, and $\boldsymbol{b}$ denotes the $N \times 1$ measurement observation vector.

4.1 Richardson iterative method

Observing that the MMSE filtering matrix $\boldsymbol{W}$ is symmetric positive definite, we can exploit the Richardson method in [17] to efficiently solve equation (4) in an iterative manner to avoid the complicated matrix inversion. The Richardson iteration can be formulated as

$$\hat{\boldsymbol{x}}^{(t+1)} = \hat{\boldsymbol{x}}^{(t)} + \eta_{\mathrm{RI}}\left(\boldsymbol{b} - \boldsymbol{A}\hat{\boldsymbol{x}}^{(t)}\right), \tag{21}$$

where the superscript t represents the number of iterations, and η_{RI} denotes the relaxation parameter of Richardson method.

According to the above description, we can use the Richardson method to estimate the transmitted signal vector $\hat{\boldsymbol{s}}$, without resorting to matrix inversion at all, as below

$$\hat{\boldsymbol{s}}^{(t+1)} = \hat{\boldsymbol{s}}^{(t)} + \eta_{\mathrm{RI}}\left(\boldsymbol{y}_{\mathrm{MF}} - \boldsymbol{W}\hat{\boldsymbol{s}}^{(t)}\right), \tag{22}$$

where $\hat{\boldsymbol{s}}^{(0)}$ denotes the initial solution, which may be dwelled on later in Section IV. The relaxation parameter η_{RI} imposes a strong impact on the convergence of Richardson method, and it always satisfies $0 < \eta_{\mathrm{RI}} < 2/\lambda_1$, where λ_1 is the largest eigenvalue of the symmetric positive definite matrix $\boldsymbol{W}$.

In accordance with (22), it can be clearly seen that the t-th iteration of the Richardson method requires one multiplication of a $2K \times 2K$ matrix $\boldsymbol{W}$ and a $2K \times 1$ vector $\hat{\boldsymbol{s}}^{(t)}$, as well as one multiplication of a constant relaxation parameter ω and a $2K \times 1$ vector $(\boldsymbol{y}_{\mathrm{MF}} - \boldsymbol{W}\hat{\boldsymbol{s}}^{(t)})$, and

thus the total computational complexity is $4K^2+2K$ for each iteration.

4.2 Successive over-relaxation method

Similar to the Richardson method, the successive over-relaxation (SOR) method is also efficient to solve the linear equation in an inversion-less way. It greatly helps one avoid the complicated matrix inversion and it is entirely different from the traditional approaches that directly computes $\boldsymbol{A}^{-1}\boldsymbol{b}$ to obtain $\boldsymbol{x}$ in equation (20). Since the matrix $\boldsymbol{A}$ is symmetric positive definite, we can decompose it into a diagonal matrix $\boldsymbol{D}_{\mathrm{A}}$, a strictly lower triangular matrix $\boldsymbol{L}_{\mathrm{A}}$, and a strictly upper triangular matrix $\boldsymbol{L}_A^H$. Then the SOR method can be represented by

$$\hat{\boldsymbol{x}}^{(t+1)}=\left(\boldsymbol{L}_{\mathrm{A}}+\frac{1}{\eta_{\mathrm{SOR}}}\boldsymbol{D}_{\mathrm{A}}\right)^{-1}\left[\left((\frac{1}{\eta_{\mathrm{SOR}}}-1)\boldsymbol{D}_{\mathrm{A}}-\boldsymbol{L}_{\mathrm{A}}^H\right)\hat{\boldsymbol{x}}^{(t)}+\boldsymbol{b}\right], \tag{23}$$

where the superscript t stands for the number of iterations, and η_{SOR} represents the relaxation parameter of SOR method, which plays an important role in convergence speed. It is noteworthy that the SOR method is equivalent to the well known Gauss-Seidel method [25] when $\omega=1$, which implies that the Gauss-Seidel method is a special case of the SOR method.

Applying the SOR method mentioned above, $\boldsymbol{W}$ can be decomposed as

$$\boldsymbol{W}=\boldsymbol{D}+\boldsymbol{L}+\boldsymbol{L}^H, \tag{24}$$

where $\boldsymbol{D}$, $\boldsymbol{L}$, and $\boldsymbol{L}^H$ denote the diagonal matrix, the strictly lower triangular matrix, and the strictly upper triangular matrix of $\boldsymbol{W}$, respectively. Thus, the transmitted signal vector in the t-th iteration is

$$\hat{\boldsymbol{s}}^{(t+1)}=\left(\boldsymbol{L}+\frac{1}{\eta_{\mathrm{SOR}}}\boldsymbol{D}\right)^{-1}\left[\left((\frac{1}{\eta_{\mathrm{SOR}}}-1)\boldsymbol{D}-\boldsymbol{L}^H\right)\hat{\boldsymbol{s}}^{(t)}+\boldsymbol{y}_{\mathrm{MF}}\right], \tag{25}$$

where $\hat{\boldsymbol{s}}^{(0)}$ denotes the initial solution, and it is set as a $2K\times1$ zero vector in generality.

Consequently, the solution to the signal problem in equation (4) can be solved by SOR method according to

$$\left(\boldsymbol{L}+\frac{1}{\eta_{\mathrm{SOR}}}\boldsymbol{D}\right)\hat{\boldsymbol{s}}^{(t+1)}=\boldsymbol{y}_{\mathrm{MF}}+\left(\left(\frac{1}{\eta_{\mathrm{SOR}}}-1\right)\boldsymbol{D}-\boldsymbol{L}^H\right)\hat{\boldsymbol{s}}^{(t)}. \tag{26}$$

Due to the fact that $(\boldsymbol{L}+1/\eta_{\mathrm{SOR}}\boldsymbol{D})$ is a lower triangular matrix, $\boldsymbol{s}^{(t+1)}$ can be obtained by solving equation (26) with low complexity. It can be found that the complexity of the t-th iteration of SOR method comes from the computation in equation (26), whose solution can be denoted as

$$\hat{s}_m^{(t+1)} = \left(1 - \frac{1}{\eta_{\mathrm{SOR}}}\right)\hat{s}_m^{(t)} + \frac{\eta_{\mathrm{SOR}}}{W_{m,m}}\left(y_{\mathrm{MF}}^m - \sum_{k<m} W_{m,k}\hat{s}_k^{(t+1)} - \sum_{k>m} W_{m,k}\hat{s}_k^{(t)}\right), \tag{27}$$

where $\hat{s}_m^{(t+1)}$, $\hat{s}_m^{(t)}$, and y_{MF}^m represent the m-th element of, $\hat{\boldsymbol{s}}^{(t)}$, and $\boldsymbol{y}_{\mathrm{MF}}$, respectively, and $W_{m,k}$ represents the (m,k) -th entry of $\boldsymbol{W}$. It is obvious that the required number of multiplications in computing $(1-\eta_{\mathrm{SOR}})\hat{\boldsymbol{s}}_m^{(t)}$ and $\frac{\eta_{\mathrm{SOR}}}{W_{m,m}}(\boldsymbol{y}_{\mathrm{MF}}^m - \sum_{k<m} W_{m,k}s_k^{(t+1)}\hat{\boldsymbol{s}}_m^{(t+1)} - \sum_{m>k} W_{m,k}s_k^{(t)}\hat{\boldsymbol{s}}_m^{(t)})$ are 1 and $2K^2+1$, respectively. Since there are $2K$ elements in $\boldsymbol{s}^{(t+1)}$ and the overall required number of multiplication is $4K^2+4K$ for each iteration.

4.3 Symmetric successive over-relaxation method

The SOR method works effectively in lowering the complexity with satisfactory performance. However, when we encounter more practically complex problems, highly tranglesome eigenvalue needs to be analyzed. Therefore, it was proposed in [19] to employ Chebyshev acceleration and symmetric successive over-relaxation (SSOR) approach to solve the problem. The SSOR method is a symmetry version of the SOR method, whose basic idea is to combine SOR with the inversed order SOR as one unified method. If, as usual, the matrix is decomposed as $\boldsymbol{W} = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{L}^H$, the SSOR method for solving the equation $\boldsymbol{W}\hat{\boldsymbol{s}} = \boldsymbol{y}_{\mathrm{MF}}$ can be carried out in the following two steps:

$Step\ 1$: Compute the previous half iteration which is identical with the SOR iteration in [18] by

$$\left(\boldsymbol{D} + \eta_{\mathrm{SSOR}}\boldsymbol{L}\right)\hat{\boldsymbol{s}}^{(t+1/2)} = (1-\eta_{\mathrm{SSOR}})\boldsymbol{D}\hat{\boldsymbol{s}}^{(t)} - \eta_{\mathrm{SSOR}}\boldsymbol{L}^H\hat{\boldsymbol{s}}^{(t)} + \eta_{\mathrm{SSOR}}\boldsymbol{y}_{\mathrm{MF}}, \tag{28}$$

$Step\ 2$: Compute the latter half iteration which is the SOR method with the equations taken in reverse order by

$$\left(\boldsymbol{D} + \eta_{\mathrm{SSOR}}\boldsymbol{L}^H\right)\hat{\boldsymbol{s}}^{(t+1)} = (1-\eta_{\mathrm{SSOR}})\boldsymbol{D}\hat{\boldsymbol{s}}^{(t+1/2)} - \eta_{\mathrm{SSOR}}\boldsymbol{L}^H\hat{\boldsymbol{s}}^{(t+1/2)} + \eta_{\mathrm{SSOR}}\boldsymbol{y}_{\mathrm{MF}}, \tag{29}$$

where $\hat{\boldsymbol{s}}^{(t)}$ represents the vector that needs to be estimated in the t-th iteration of SSOR method, and $\hat{\boldsymbol{s}}^{(0)}$ denotes the initial solution of SSOR, which is usually chosen as a $2K\times 1$ zero vector without loss of generality.

Compared with SOR, the SSOR method has two advantages. Firstly, the structure of SSOR method is symmetric, which implies that the convergence rate of SSOR can be improved by using Chebyshev acceleration. Secondly, a simple and quantified relaxation parameter can be employed to approximate a precise relaxation parameter with negligible performance loss, when we consider the convergence rate of SSOR method is not very sensitive to the relaxation parameter η_{SSOR} . A detailed description of the relaxation parameters is given in the subsequent subsection.

On the basis of equation (28) and (29), it is obvious that the computation of each element of $\hat{\boldsymbol{s}}^{(t+1/2)}$ and $\hat{\boldsymbol{s}}^{(t+1)}$ requires $K+1$ times of multiplications. Since there are K elements in them, the required number of multiplications of SSOR is $t(8K^2+8K)$.

Based on the idea of iterative SOR and SSOR, we are able to employ an averaging weight to deal with the vector derived by the iteration in equation (25) and the vector derived by iteration in equation (29). The weighted symmetric successive over-relaxation (WSSOR) method can be described as

$$\hat{\boldsymbol{s}}^{(t+1)} = \boldsymbol{B}\hat{\boldsymbol{s}}^{(t)} + \boldsymbol{C}, \tag{30}$$

where

$$\boldsymbol{B} = \theta(1-\eta_{\mathrm{WSSOR}})((\boldsymbol{D}+\eta_{\mathrm{WSSOR}}\boldsymbol{L})^{-1} - (\boldsymbol{D}+\eta_{\mathrm{WSSOR}}\boldsymbol{L}^H))^{-1}\boldsymbol{D} + \theta\eta_{\mathrm{WSSOR}}((\boldsymbol{D}+\eta_{\mathrm{WSSOR}}\boldsymbol{L})^{-1})\boldsymbol{L}^H - (\boldsymbol{D}+\eta_{\mathrm{WSSOR}}\boldsymbol{L}^H)^{-1}\boldsymbol{L}$$
$$+(1-\eta_{\mathrm{WSSOR}})(\boldsymbol{D}+\eta_{\mathrm{WSSOR}}\boldsymbol{L}^H)^{-1}\boldsymbol{D} + \eta_{\mathrm{WSSOR}}(\boldsymbol{D}+\eta_{\mathrm{WSSOR}}\boldsymbol{L}^H)^{-1}\boldsymbol{L},$$
$$\boldsymbol{C} = \eta_{\mathrm{WSSOR}}(\boldsymbol{D}+\eta_{\mathrm{WSSOR}}\boldsymbol{L})\boldsymbol{y}_{\mathrm{MF}},$$

and the weighting factor is $\theta \in [0,1]$.

4.4 Conjugate gradient algorithm

In this subsection, based on the significant channel characteristic exhibited in massive MIMO, conjugate gradient (CG) method was employed in [20] to iteratively achieve the MMSE estimate without matrix inversion and the transmitted signal vector can be obtained as

$$\hat{\boldsymbol{s}}^{(t+1)} = \hat{\boldsymbol{s}}^{(t)} + \delta^{(t)}\boldsymbol{p}^{(t)}, \tag{31}$$

where $\hat{\boldsymbol{s}}^{(0)}$ is the initial vector of CG method, which is usually selected as a zero vector, $\delta^{(t)}$ is a scalar parameter, and $\boldsymbol{p}^{(t)}$ is the conjugate direction with regard to $\mathbf{W}$, i.e.,

$$(\boldsymbol{p}^{(t)})^H\boldsymbol{W}\boldsymbol{p}^{(j)}, \quad for \quad t \neq j \tag{32}$$

Let $\boldsymbol{r}^{(t)}$ represent the estimate error residual in the t-th iteration, described as

$$\boldsymbol{r}^{(t)} = \boldsymbol{y}_{\mathrm{MF}} - \boldsymbol{W}\hat{\boldsymbol{s}}^{(t)}. \tag{33}$$

By reason that $\boldsymbol{p}^{(t)}$ is the conjugate direction with respect to $\boldsymbol{W}$, it can be represented as

$$\boldsymbol{p}^{(t)} = \boldsymbol{r}^{(t)} - \sum_{k<t}\frac{(\boldsymbol{p}^{(k)}, \boldsymbol{W}\boldsymbol{r}^{(t)})}{(\boldsymbol{p}^{(k)}, \boldsymbol{W}\boldsymbol{p}^{(k)})}\boldsymbol{p}^{(k)}. \tag{34}$$

Provided that we have

$$\begin{aligned}\boldsymbol{r}^{(t)} &= \boldsymbol{y}_{\mathrm{MF}} - \boldsymbol{W}\hat{\boldsymbol{s}}^{(t)} \\ &= \boldsymbol{y}_{\mathrm{MF}} - \boldsymbol{W}\left(\hat{\boldsymbol{s}}^{(t-1)} + \delta^{(t-1)}\boldsymbol{p}^{(t-1)}\right) \\ &= \left(\boldsymbol{y}_{\mathrm{MF}} - \boldsymbol{W}\hat{\boldsymbol{s}}^{t-1}\right) - \delta^{(t-1)}\boldsymbol{W}\boldsymbol{p}^{(t-1)} \\ &= \boldsymbol{r}^{(t-1)} - \delta^{(t-1)}\boldsymbol{W}\boldsymbol{p}^{(t-1)},\end{aligned} \tag{35}$$

we can derive from the above formula

$$\boldsymbol{W}\boldsymbol{p}^{(t-1)} = \frac{\boldsymbol{r}^{(t-1)} - \boldsymbol{r}^{(t)}}{\delta^{(t-1)}}. \tag{36}$$

Since $\boldsymbol{W}$ is Hermitian positive definite, we have

$$\begin{aligned} \left(\boldsymbol{W}\boldsymbol{r}^{(t)}, \boldsymbol{p}^{t}\right) &= \left(\boldsymbol{r}^{(t)}, \boldsymbol{W}\boldsymbol{p}^{(t)}\right) \\ &= \left(\boldsymbol{r}^{(t)}, \frac{\boldsymbol{r}^{(j)} - \boldsymbol{r}^{(j+1)}}{\delta^{(j)}}\right) \\ &= \frac{1}{\delta^{(j)}}\left[\left(\boldsymbol{r}^{(t)}, \boldsymbol{r}^{(j)}\right) - \left(\boldsymbol{r}^{(t)}, \boldsymbol{r}^{(j+1)}\right)\right]. \end{aligned} \tag{37}$$

In accordance with [26, Theorem 10.2.3], we have

$$\left(\boldsymbol{r}^{(t)}, \boldsymbol{r}^{(j)}\right) = 0, \quad \text{if} \quad t \neq j, \tag{38}$$

then, (37) can be rewritten as

$$\left(\boldsymbol{W}\boldsymbol{r}^{(t)}, \boldsymbol{p}^{(j)}\right) = 0, \quad \text{if} \quad j < t-1. \tag{39}$$

Hence, by substituting (39) into (34), $\boldsymbol{p}^{(t)}$ can be further deduced as

$$\boldsymbol{p}^{(t)} = \boldsymbol{r}^{(t)} - \frac{\left(\boldsymbol{p}^{(t-1)}, \boldsymbol{W}\boldsymbol{r}^{(t)}\right)}{\left(\boldsymbol{p}^{(t-1)}, \boldsymbol{W}\boldsymbol{p}^{(t-1)}\right)}. \tag{40}$$

If we define

$$\zeta^{(t)} \triangleq -\frac{\left(\boldsymbol{p}^{(t)}, \boldsymbol{W}\boldsymbol{r}^{(t+1)}\right)}{\left(\boldsymbol{p}^{(t)}, \boldsymbol{W}\boldsymbol{p}^{(t)}\right)} = \frac{\left(\boldsymbol{r}^{(t+1)}, \boldsymbol{W}\boldsymbol{p}^{(t)}\right)}{\left(\boldsymbol{p}^{(t)}, \boldsymbol{W}\boldsymbol{p}^{(t-1)}\right)}, \tag{41}$$

then, equation (40) can be rewritten as

$$\boldsymbol{p}^{(t)} = \boldsymbol{r}^{(t)} + \zeta^{(t-1)}\boldsymbol{p}^{(t-1)}, \tag{42}$$

where $\boldsymbol{p}^{(0)}$ stands for the initial search direction which is usually selected to be $\boldsymbol{r}^{(0)}$ in general case.

Aiming to reduce the computational complexity, we can choose $\delta^{(t)}$ as

$$\delta^{(t)} = \frac{\left(\boldsymbol{r}^{(t)}, \boldsymbol{r}^{(t)}\right)}{\left(\boldsymbol{r}^{(t)}, \boldsymbol{W}\boldsymbol{p}^{(t)}\right)}. \tag{43}$$

From the above analysis, we can conclude that the computational complexity in the t-th iteration consists of computing $\delta^{(t)}$, $\boldsymbol{r}^{(t+1)}$, $\hat{\boldsymbol{s}}^{(t+1)}$, $\boldsymbol{p}^{(t+1)}$ and $\zeta^{(t)}$. Computing $\delta^{(t)}$ and $\boldsymbol{r}^{(t+1)}$ involves $2K^2+4K$ and $2K^2+2K$, respectively. In addition, $2K$ times of multiplication are required to compute $\hat{\boldsymbol{s}}^{(t+1)}$ and $\boldsymbol{p}^{(t+1)}$, respectively. Calculating $\zeta^{(t)}$ requires $4K$ times of multiplication. To sum up, totally $4K^2+14K$ times of multiplication are required for each iteration of CG method.

4.5 Simplified approximate message passing method

It is noteworthy that *a priori* knowledge of the distribution of different input signals should be taken into account in the derivation process when conventional AMP algorithm is used to solve the LASSO problem in the field of compressive sensing, and consequently the algorithm may be limited by the distribution of the input signal. As for the MMSE detector, it does not consider the exact a priori information of statistical distribution of the received signal, which is usually assumed to be Gaussian. Hence, utilizing the SAMP method to estimate the transmitted signal vector $\hat{\boldsymbol{s}}$ can be presented by

$$\hat{\boldsymbol{s}}^{(t)}=\gamma^{(t)}\left(\boldsymbol{H}^H\boldsymbol{z}^{(t)}+\hat{\boldsymbol{s}}^{(t-1)}\right), \tag{44}$$

$$\boldsymbol{z}^{(t)}=\boldsymbol{y}-\boldsymbol{H}\hat{\boldsymbol{s}}^{(t-1)}+\beta\gamma^{(t-1)}\boldsymbol{r}^{(t-1)}, \tag{45}$$

where system configuration parameter $\beta=K/N$, $\hat{\boldsymbol{s}}^{(t)}$ denotes the estimated vector in the t-th iteration of SAMP method, and $\boldsymbol{z}^{(t)}$ is the message passing parameter in the t -th iteration. The message passing term $\beta\gamma^{(t-1)}\boldsymbol{r}^{(t-1)}$ approximates the combined effect on the reconstruction of the passing NK messages in the full message passing algorithm. The coefficient $\gamma^{(t)}$ denotes the soft thresholding value in the t-th iteration, which determines the convergence speed by constructing a similar coefficient form of signal-to-noise ratio that can be expressed as

$$\gamma^{(t)}=\frac{\sigma_s^2}{\sigma_s^2+\alpha^{(t)}}, \tag{46}$$

where $\alpha^{(t)}$ denotes the effective noise variance in the t-th iteration of SAMP method, which can effectively control the soft threshold effect of $\gamma^{(t)}$ in equation (44), and improve the detection performance and convergence rate of SAMP method.

The SAMP method we propose in this paper is very attractive because it requires neither iteratively updating the mean and variance nor approximating the complicated matrix inversion for solving the linear equation. It just demands simple operations on the matrix $\boldsymbol{H}$ and $\boldsymbol{H}^H$. In terms of complexity, the SAMP method only costs $\mathcal{O}(NK)$ in each iteration. In contrast, it is worth reminding that computing the MMSE exact matrix inversion requires a cubic computation complexity $\mathcal{O}(K^3)$.

The computational complexity and performance of all the aforementioned simplified matrix inversion algorithms are compared and analyzed in **Table 1**. Since both the conventional MMSE algorithm and its common simplified algorithms need to compute $\boldsymbol{W}$ and $\boldsymbol{y}_{\mathrm{MF}}$, computation analysis is not given for obtaining them, and thus an emphasis is made on

discussing the iteration process, and evaluating it in terms of the required number of multiplications in acquiring the simplified matrix inversion.

Table 1. The comparative analysis of detection performance and computational complexity

ALGORITHM	CORE FORMULA	FORMULA NUMBER	PERFORMANCE (In descending order)	COMPLEXITY
MMSE	$\boldsymbol{W}=\boldsymbol{G}+\sigma^2\boldsymbol{I}_{2K}$,	(5)	OPTIMAL	$\mathcal{O}(K^3)$
Simplified approximate message passing [22](SAMP)	$\hat{\boldsymbol{s}}^{(t)}=\gamma^{(t)}\left(\boldsymbol{H}^H\boldsymbol{z}^{(t)}+\hat{\boldsymbol{s}}^{(t-1)}\right)$, $\boldsymbol{z}^{(t)}=\boldsymbol{y}-\boldsymbol{H}\hat{\boldsymbol{s}}^{(t-1)}+\beta\gamma^{(t-1)}\boldsymbol{r}^{(t-1)}$,	(44) (45)	SUB-OPTIMAL	$\mathcal{O}(NK)$
Multistage linear receiver [15](MLR)	$\boldsymbol{W}^{-1}=(\boldsymbol{G}+\sigma^2\boldsymbol{I}_{2K})^{-1}$ $=\sum_{s=0}^{t}\omega_s(\boldsymbol{G})^s$	(11)	SUB-OPTIMAL	$\mathcal{O}(K^2)$
symmetric successive over-relaxation [19](SSOR)	$(\boldsymbol{D}+\eta_{\mathrm{SSOR}}\boldsymbol{L})\hat{\boldsymbol{s}}^{(t+1/2)}=(1-\eta_{\mathrm{SSOR}})\boldsymbol{D}\hat{\boldsymbol{s}}^{(t)}$ $-\eta_{\mathrm{SSOR}}\boldsymbol{L}^H\hat{\boldsymbol{s}}^{(t)}+\eta_{\mathrm{SSOR}}\boldsymbol{y}_{\mathrm{MF}}$ $(\boldsymbol{D}+\eta_{\mathrm{SSOR}}\boldsymbol{L}^H)\hat{\boldsymbol{s}}^{(t+1)}=(1-\eta_{\mathrm{SSOR}})\boldsymbol{D}\hat{\boldsymbol{s}}^{(t+1/2)}$ $-\eta_{\mathrm{SSOR}}\boldsymbol{L}^H\hat{\boldsymbol{s}}^{(t+1/2)}+\eta_{\mathrm{SSOR}}\boldsymbol{y}_{\mathrm{MF}}$	(28) (29)	SUB-OPTIMAL	$\mathcal{O}(K^2)$
Successive over-relaxation [18](SOR)	$\hat{\boldsymbol{s}}^{(t+1)}=(\boldsymbol{L}+\frac{1}{\eta_{\mathrm{SOR}}}\boldsymbol{D})^{-1}\left[\left(\left(\frac{1}{\eta_{\mathrm{SOR}}}-1\right)\boldsymbol{D}-\boldsymbol{L}^H\right)\hat{\boldsymbol{s}}^{(t)}+\boldsymbol{y}_{\mathrm{MF}}\right]$	(23)	BETTER	$\mathcal{O}(K^2)$
Newton Iterative Approximation [16](NT)	$\boldsymbol{W}_t^{-1}=\boldsymbol{W}_{t-1}^{-1}\left(2\boldsymbol{I}_{2K}-\boldsymbol{W}\boldsymbol{W}_{t-1}^{-1}\right)$	(17)	BETTER	$\mathcal{O}(K^2)$
Conjugate gradient [20](CG)	$\hat{\boldsymbol{s}}^{(t+1)}=\hat{\boldsymbol{s}}^{(t)}+\delta^{(t)}\boldsymbol{p}^{(t)}$ $\boldsymbol{p}^{(t)}=\boldsymbol{r}^{(t)}-\sum_{k<t}\frac{\left(\boldsymbol{p}^{(k)},\boldsymbol{W}\boldsymbol{r}^{(t)}\right)}{\left(\boldsymbol{p}^{(k)},\boldsymbol{W}\boldsymbol{p}^{(k)}\right)}\boldsymbol{p}^{(k)}$ $\delta^{(t)}=\frac{\left(\boldsymbol{r}^{(t)},\boldsymbol{r}^{(t)}\right)}{\left(\boldsymbol{r}^{(t)},\boldsymbol{W}\boldsymbol{p}^{(t)}\right)}$	(31) (34) (43)	BETTER	$\mathcal{O}(K^2)$
Richardson iterative [17](RI)	$\hat{\boldsymbol{s}}^{(t+1)}=\hat{\boldsymbol{s}}^{(t)}+\eta_{\mathrm{RI}}\left(\boldsymbol{y}_{\mathrm{MF}}-\boldsymbol{W}\hat{\boldsymbol{s}}^{(t)}\right)$	(22)	GOOD	$\mathcal{O}(K^2)$
Neumann Series Expansion [13](NI)	$\boldsymbol{W}^{-1}=\sum_{n=0}^{t-1}(-\boldsymbol{D}^{-1}\boldsymbol{E})^n\boldsymbol{D}^{-1}$	(8)	NORMAL	$t=1,\mathcal{O}(K)$ $t=2,\mathcal{O}(K^2)$ $t=3,\mathcal{O}(K^3)$
Diagona Approximation	$\boldsymbol{W}=\boldsymbol{D}^H\boldsymbol{D}+\sigma^2\boldsymbol{I}_{2K}$	(6)	WORST	$\mathcal{O}(K)$

The computation complexity comparison among all simplified algorithms is shown in **Fig.1**. It is clear that the computational complexity of all the simplified algorithms increases when the number of users and number of iterations or expansion terms is increased. Unlike the Neumann series expansion and the Newton iterative method with the number of expansion terms exceeding three($t\geq 3$) or two($t\geq 2$), whose complexity is surpassing that of the exact matrix inversion, the MLR method and the methods based on linear equation solving can reduce it from $\mathcal{O}(K^3)$ to $\mathcal{O}(K^2)$ for any arbitrary number of iterations. Compared with other

methods whose complexity is, the SSOR and SAMP method have the highest and the least complexity for the same iteration respectively, and the MLR, CG, RI and SOR method have approximately the equal complexity. In the meantime, the complexity of the SAMP method exceeds other methods based on linear equation solving with the same number of iterations, when the number of users is small. However, its complexity is asymptotically inferior to other simplified algorithms when the number of users is increased.

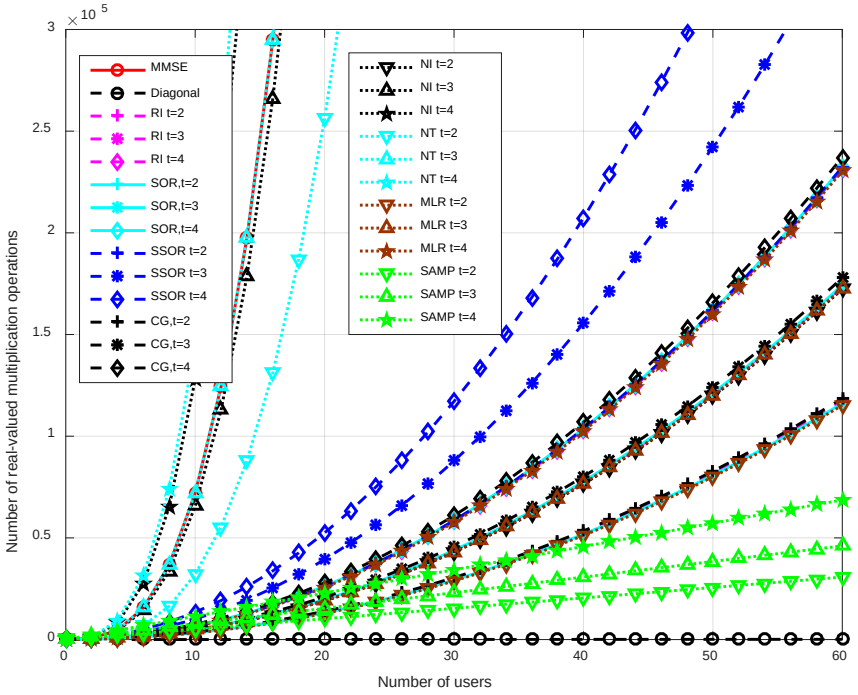

Fig. 1. computation complexity comparison among all simplified algorithms

5. Performance optimizations

In order to further improve the performance under the premise of low complexity and in the mean time ensure the feasibility of the simplified algorithms in practice, we propose some optimization schemes in terms of relaxation parameter, initial solution, and effective noise variance.

5.1 The quantified relaxation parameter

As can be seen from equation (22), (23) and (29), the selection of relaxation parameter will pose effect on the convergence rate of the proposed algorithms which are based on solving the given linear equation. It has been proved in [27] that the optimal relaxation parameter η^{opt} can be given by

$$\eta^{opt} = \frac{2}{1+\sqrt{2(1-\rho(\boldsymbol{B}_J))}}, \tag{47}$$

where $\rho(\boldsymbol{B}_J)$ is the spectral radius of Jacobi iteration matrix $\boldsymbol{B}_J$, which can be expressed as

$$\boldsymbol{B}_J = \boldsymbol{D}^{-1}\boldsymbol{W} - \boldsymbol{I}_{2\mathrm{K}}. \tag{48}$$

Each element of the diagonal component $\boldsymbol{D}$ will tends towards to a fixed value N in massive MIMO systems[5], which indicates that

$$\boldsymbol{D}^{-1} \approx \frac{1}{N}\boldsymbol{I}_{2\mathrm{K}}. \tag{49}$$

Moreover, as the matrix $\boldsymbol{W}$ is a central Wishart matrix, when the number of N and K is sufficiently large and the system configuration ratio $\beta = K/N$ remains fixed, the largest eigenvalue λ_{max} of $\boldsymbol{W}$ can be well approximated by [5]

$$\lambda_{max} = N\left(1+\sqrt{\beta}\right)^2. \tag{50}$$

Actually, the relaxation parameter depends more on the average eigenvalue of $\boldsymbol{W}$ than the largest eigenvalue. However, to obtain the average eigenvalue we must compute all the eigenvalues of $\boldsymbol{W}$ within the symbol period or frame duration, resulting in highly burdensome complexity in practice. Therefore, we can propose a simple proper relaxation parameter $\overline{\eta}$ to replace η^{opt} in equation (47) with negligible error as

$$\overline{\eta} = \frac{2}{1+\sqrt{2(1-a)}}, \quad a = \left(1+\sqrt{\beta}\right)^2 - 1, \tag{51}$$

where it means that the proper relaxation parameter $\overline{\eta}$ is merely determined by the system configuration ratio β. In practice, $\overline{\eta}$ will usually be a constant when the realistic antenna configuration of massive MIMO system is fixed. Thus the proposed proper relaxation parameter $\overline{\eta}$ can be expected to simplify the implementation and maintain the performance of signal detection.

5.2 The proper initial solution

In order to facilitate implementation, the traditional iterative algorithms usually set the initial solution as an all-zero vector. However, choosing a suitable initial solution can speed up the convergence rate. Moreover, better performance can be achieved by choosing a proper initial solution than the all-zero vector under the same number of iterations.

When β is large enough, $\boldsymbol{D}^{-1}$ is a very good approximation for $\boldsymbol{W}^{-1}$, and $\boldsymbol{G} \approx N\boldsymbol{I}_{2\mathrm{K}}$ is in accordance with the channel hardening phenomenon [28], where the off-diagonal entries of $\mathbf{G}$ become increasingly weaker compared to the diagonal ones, we can obtain $\boldsymbol{W}^{-1} \approx \boldsymbol{D}^{-1} \approx N^{-1}\boldsymbol{I}_{2\mathrm{K}}$. Afterwards, the proper initial solutions in equation (22), (25), (29), (31), and (44) can be generally set as

$$\hat{\boldsymbol{s}}^{(0)} = \frac{1}{N}\boldsymbol{I}_{2K}\boldsymbol{y}_{\mathrm{MF}}. \tag{52}$$

5.3 The effective noise variance

According to equation (44) and (45) we can analyze the convergence rate and optimize the performance of SAMP method through defining the effective variance and setting the corresponding parameters. The effective variance is composed of the Gaussian white noise variance and the additional interference term [29] and it is expressed as

$$\alpha^{(t)} = \sigma_z^2 + \beta\Phi(\alpha^{(t-1)}), \tag{53}$$

where $\beta\Phi(\alpha^{(t-1)})$ approximates the additional interference term, indicating the interference from other antennas. According to the work in [30], we have the following upper bound of mean squared error function

$$\Phi(\alpha^{(t-1)}) \le \frac{\sigma_s^2\alpha^{(t-1)}}{\sigma_s^2 + \alpha^{(t-1)}}. \tag{54}$$

If the upper bound of the MSE function is chosen, equation (53) can be rewritten as

$$\alpha^{(t)} = \sigma_z^2 + \beta\frac{\sigma_s^2\alpha^{(t-1)}}{\sigma_s^2 + \alpha^{(t-1)}}. \tag{55}$$

In general cases, the effective variance will be initialized to $\alpha^{(0)} = \sigma_z^2$. However, this rough approximation completely ignores the interference term from other antennas. To achieve the exact initial effective variance, an additional term was added for compensation, which can be expressed as

$$\alpha^{(0)} = \sigma_z^2 + \beta Var[\boldsymbol{s}] = \sigma_z^2 + \sigma_s^2 \ . \tag{56}$$

6. Simulation Result

To verify the performance of the simplified algorithms, we provide the uncoded bit error rate (BER) simulation results in this section. The BER performance of the MMSE algorithm with exact matrix inversion and its most simplified version of diagonal matrix inversion are included as the benchmarks for comparison. We consider the modulation scheme of 16QAM, and the massive MIMO system configuration is $N \times K = 128 \times 16$. In the following simulation diagrams, the parameter t denotes the number of iterations for the simplified algorithm based on linear equation solving, and the number of terms adopted in the matrix inversion approximation based algorithms as well. We assume that perfect knowledge of $\boldsymbol{H}$ is readily known at the receiver with uncorrelated Rayleigh fading channel effects in simulations.

Fig. 2 shows the BER performance comparison among the simplified algorithms based on matrix inversion approximation. It is observed that the BER performance of all the algorithms described in Section III improves when the number of iterations is increased. In the category of

matrix inversion approximation based algorithms, the MLR method outperforms the Neumann series expansion and the Newton iterative method with the same number of iterations, whereas the Newton iterative method is superior to the Neumann series expansion approach. For example, when $t = 3$, to achieve the BER performance of 10^{-5}, the required SNR for the MLR method is just 11dB, while the Newton iterative approximation required 12dB and the Neumann series expansion demands more than 16dB. Furthermore, we can make a conclusion from the simulation results that the MLR method yields the BER performance near that of the MMSE detector with only a small number of expansion terms.

Fig. 3 compares the BER performance among the algorithms based on linear equation solving with the proper relaxation parameter and the proper initial solution proposed in Section V. As shown, the BER performance of all algorithms described in Section IV improves when we increase the number of iterations. Within this algorithm category, the SAMP and SSOR methods achieve the best performance, the SOR and CG methods are inferior to them, and the Richardson method obtains the worst performance, when the number of iterations is set as $t \geq 3$. For example, when $t = 3$, to achieve the BER performance of 10^{-5}, the SNR required by the SAMP method and the SSOR method is just about 11dB, the CG and SOR method require approximately 12dB and 14dB, respectively, while the Richardson method requires more than 16dB. In the meantime, the SAMP method is slightly better compared with SOR method for the number of iterations $t = 2$.

The overall comparison of the two categories of simplified algorithms is given in **Fig. 4**. It is clear that the BER performance is ameliorated for all the algorithms with increased number of iterations or expansion terms. **Fig. 5** compares the BER performance among SAMP, MLR, and SSOR algorithms. It is clear that the performances of them are more prominent among the simplified algorithms and their performances are quite close to that of the MMSE detector with exact matrix inversion, when the number of iteration $t = 2$. More specifically, the performance of the SAMP method outperforms MLR and SSOR with the SNR gain of 0.5dB and 2dB respectively, when the number of iterations is set as $t = 2$ to achieve BER of 10^{-5}.

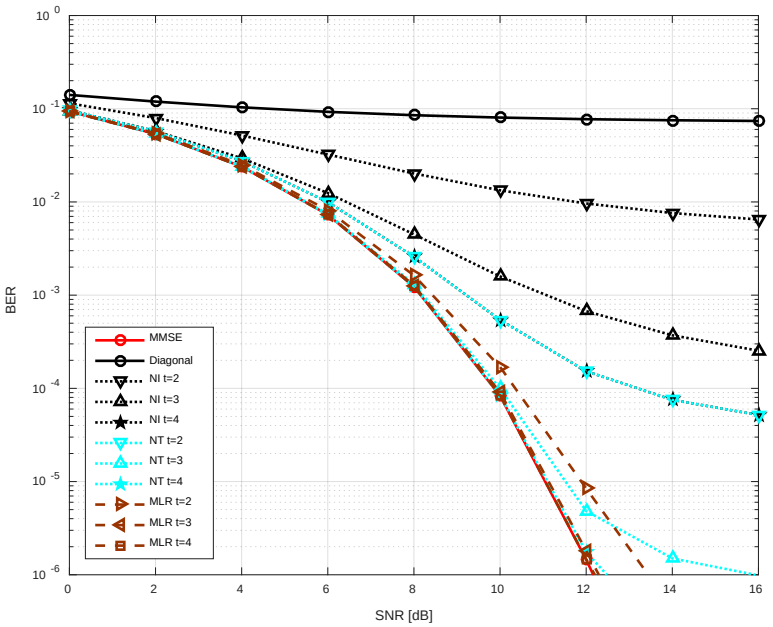

Fig. 2. BER performance comparison between the algorithms based on matrix inversion reduction

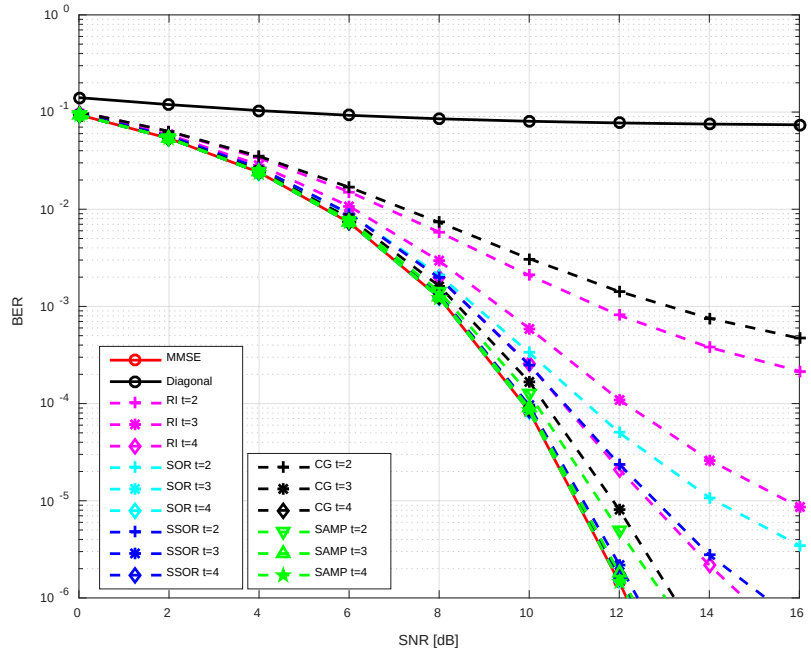

Fig. 3. BER performance comparison between the algorithms based on linear equation solving

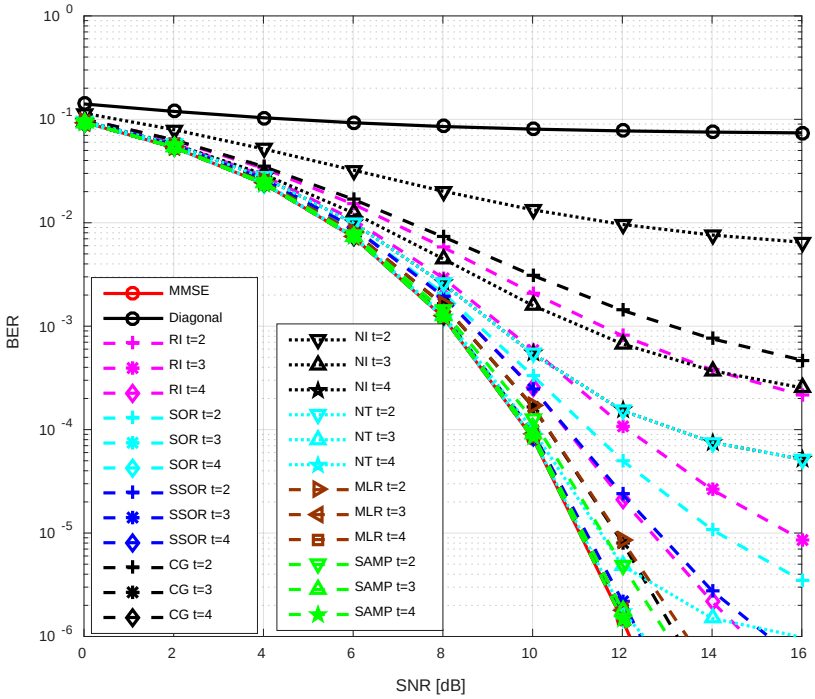

Fig. 4. BER performance comparison among all simplified algorithms

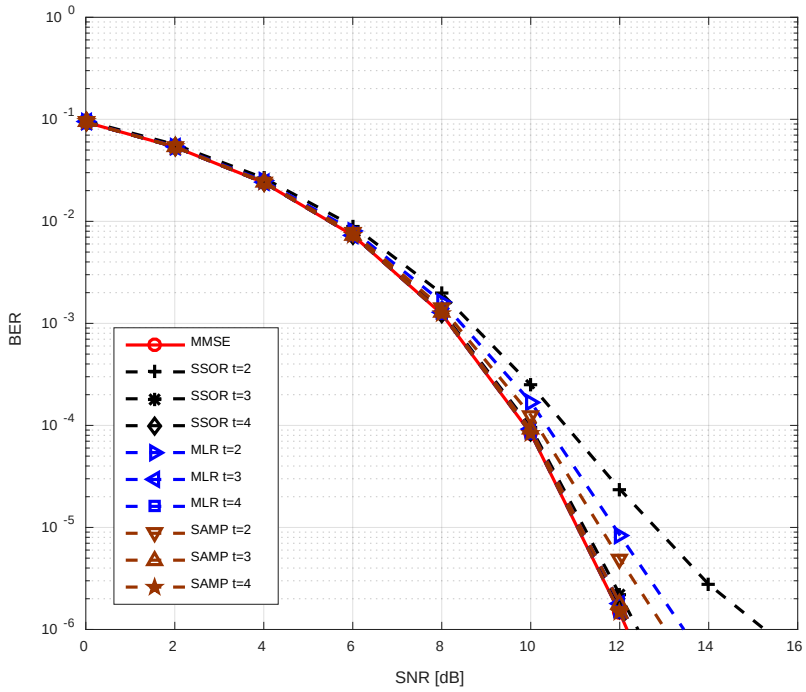

Fig. 5. BER performance comparison among SSOR, MLR and SAMP

7. Conclusion

In this paper, we compare and analyze the simplified detection algorithms to avoid the complicated matrix inversion that is indispensable for the conventional MMSE detector. These algorithms can reduce the computational complexity from $\mathcal{O}(K^3)$ to $\mathcal{O}(K^2)$ or $\mathcal{O}(NK)$ by setting the number of expansion terms or iterations. Meanwhile, we divided these algorithms into two categories and make detailed comparison, respectively. In order to guarantee satisfactory performance in practice, a simplified AMP algorithm and some optimizing measures for setting the algorithm parameters and initial solutions are proposed, further improving the performance with a low computational complexity. Simulation results show that the MLR method performs best among the algorithms based on matrix inversion approximation and the proposed SAMP method is the most prominent one within the algorithms based on linear equation solving. The proposed SAMP method outperforms the MLR method with a slightly higher complexity and they may serves as salient candidate detection schemes for massive MIMO systems in practice.

References

[1] A. Lozano and N. Jindal, "Transmit diversity vs. spatial multiplexing in modern MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 1, pp. 186-197, January, 2010. Article (CrossRef Link)

[2] D. Gesbert, M. Kountouris, J. R. W. Heath, C. b. Chae and T. Salzer, "Shifting the MIMO Paradigm," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 36-46, September, 2007. Article (CrossRef Link)

[3] X. Ge, R. Zi, H. Wang, J. Zhang and M. Jo, "Multi-User Massive MIMO Communication Systems Based on Irregular Antenna Arrays," *IEEE Transactions on Wireless Communications*, vol. 15, no.

8, pp. 5287-5301, August, 2016. Article (CrossRef Link)
[4] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) Multiplexing and channel coding (Release 9), TS 36. 212 Rev. 8.3.0, 3GPP Organizational Partners, May, 2008. Article (CrossRef Link)
[5] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors and F. Tufvesson, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40-60, January, 2013. Article (CrossRef Link)
[6] J. Jose, A. Ashikhmin, T. L. Marzetta and S. Vishwanath, "Pilot Contamination and Precoding in Multi-Cell TDD Systems," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2640-2651, August, 2011. Article (CrossRef Link)
[7] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74-80, February, 2014. Article (CrossRef Link)
[8] M. Qian, Y. Wang, Y. Zhou and J. Shi, "A super base station based centralized network architecture for 5G mobile communication systems," *Digital Communications and Networks* vol. 1, no. 2, pp. 152-159, April, 2015. Article (CrossRef Link)
[9] D.C. Araújo, T. Maksymyuk, et al, "Massive MIMO: Survey and Future Research Topics," *IET Communications*, vol. 10, no.15, pp. 1938-1946, October, 2016. Article (CrossRef Link)
[10] N. Srinidhi, T. Datta, A. Chockalingam and B. S. Rajan, "Layered Tabu Search Algorithm for Large-MIMO Detection and a Lower Bound on ML Performance," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 2955-2963, November, 2011. Article (CrossRef Link)
[11] L. G. Barbero and J. S. Thompson, "Fixing the Complexity of the Sphere Decoder for MIMO Detection," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2131-2142, June, 2008. Article (CrossRef Link)
[12] J. Goldberger and A. Leshem, "MIMO Detection for High-Order QAM Based on a Gaussian Tree Approximation," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4973-4982, August, 2011. Article (CrossRef Link)
[13] M. Wu, B. Yin, G. Wang, C. Dick, J. R. Cavallaro and C. Studer, "Large-Scale MIMO Detection for 3GPP LTE: Algorithms and FPGA Implementations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 916-929, October, 2014. Article (CrossRef Link)
[14] B. Kang, J. H. Yoon and J. Park, "Low complexity massive MIMO detection architecture based on Neumann method," in *Proc. of 2015 International SoC Design Conference (ISOCC)* pp. 293-294, November 2-5, 2015. Article (CrossRef Link)
[15] T. Li, S. Patole and M. Torlak, "A multistage linear receiver approach for MMSE detection in massive MIMO," in *Proc. of 2014 48th Asilomar Conference on Signals, Systems and Computers*, pp. 2067-2072, November 2-5, 2014. Article (CrossRef Link)
[16] C. Tang, C. Liu, L. Yuan and Z. Xing, "High Precision Low Complexity Matrix Inversion Based on Newton Iteration for Data Detection in the Massive MIMO," *IEEE Communications Letters*, vol. 20, no. 3, pp. 490-493, March, 2016. Article (CrossRef Link)
[17] X. Gao, L. Dai, C. Yuen and Y. Zhang, "Low-complexity MMSE signal detection based on Richardson method for large-scale MIMO systems," in *Proc. of 2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, pp. 1-5, Stepmother 14-17, 2014. Article (CrossRef Link)
[18] X. Gao, L. Dai, Y. Hu, Z. Wang and Z. Wang, "Matrix inversion-less signal detection using SOR method for uplink large-scale MIMO systems," in *Proc. of 2014 IEEE Global Communications Conference*, pp. 3291-3295, December 8-12, 2014. Article (CrossRef Link)
[19] J. Ning, Z. Lu, T. Xie and J. Quan, "Low complexity signal detector based on SSOR method for massive MIMO systems," in *Proc. of 2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1-4, June 17-19, 2015. Article (CrossRef Link)
[20] Y. Hu, Z. Wang, X. Gaol and J. Ning, "Low-complexity signal detection using CG method for uplink large-scale MIMO systems," in *Proc. of 2014 IEEE International Conference on Communication Systems*, pp. 477-481, November 19-21, 2014. Article (CrossRef Link)

[21] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang and Q. Guo, "Low-Complexity Iterative Detection for Large-Scale Multiuser MIMO-OFDM Systems Using Approximate Message Passing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 902-915, October, 2014. Article (CrossRef Link)
[22] D. L. Donoho, A. Maleki and A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *Proc. of 2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, pp. 1-5, January 6-8, 2010. Article (CrossRef Link)
[23] M. Bayati and A. Montanari, "The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764-785, February, 2011. Article (CrossRef Link)
[24] B. Ren, Y. Wang, S. Sun, Y. Zhang, X. Dai and K. Niu, "Low-complexity MMSE-IRC algorithm for uplink massive MIMO systems," *Electronics Letters*, vol. 53, no. 14, pp. 972-974, July, 2017. Article (CrossRef Link)
[25] L. Dai, X. Gao, X. Su, S. Han, I. C. L and Z. Wang, "Low-Complexity Soft-Output Signal Detection Based on Gauss-Seidel Method for Uplink Multiuser Large-Scale MIMO Systems," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4839-4845, October, 2015. Article (CrossRef Link)
[26] L. Dai, Z. Wang and Z. Yang, "Time-Frequency Training OFDM with High Spectral Efficiency and Reliable Performance in High Speed Environments," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 4, pp. 695-707, May, 2012. Article (CrossRef Link)
[27] Björck, Åke. Numerical methods in matrix computations. New York: Springer, 2015. Article (CrossRef Link)
[28] T. L. Narasimhan and A. Chockalingam, "Channel Hardening-Exploiting Message Passing (CHEMP) Receiver in Large-Scale MIMO Systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 847-860, October, 2014. Article (CrossRef Link)
[29] A. Maleki, L. Anitori, Z. Yang and R. G. Baraniuk, "Asymptotic Analysis of Complex LASSO via Complex Approximate Message Passing (CAMP)," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4290-4308, July, 2013. Article (CrossRef Link)
[30] D. Guo, Y. Wu, S. S. Shitz and S. Verdu, "Estimation in Gaussian Noise: Properties of the Minimum Mean-Square Error," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2371-2385, April, 2011. Article (CrossRef Link)

Shufeng Zhao has been pursuing his M.Sc. degree at Chongqing University of Posts and Telecommunications (CQUPT) since September 2015. His current research interests include signal detection for massive MIMO uplink systems, LTE systems and digital signal processing.

Bin Shen received the M.Sc. and Ph.D. degrees in communication engineering from University of Electronic Science and Technology of China (UESTC) in 2005, and Inha University, Korea, in 2010, respectively. He is now a professor with the school of communication and information engineering, Chongqing University of Posts and Telecommunications (CQUPT). His research interests are in statistical signal processing, MIMO systems, and cognitive radios.

Quan Hua received his M.Sc. degree in Electronics and Communication Engineering from Chongqing University of Posts and Telecommunications (CQUPT) in 2017. His research interests include massive MIMO systems, 3GPP LTE systems and digital signal processing.