Iterative Matrix Inversion Based Low Complexity Detection in Large/Massive MIMO Systems

Vipul Gupta*, Abhay Kumar Sah† and A. K. Chaturvedi‡
Department of Electrical Engineering,
Indian Institute of Technology Kanpur
Kanpur, India 208016
Email: {vipgupta*, abhaysah†, akc‡}@iitk.ac.in

***Abstract*—Linear detectors such as zero forcing (ZF) or minimum mean square error (MMSE) are imperative for large/massive MIMO systems for both the downlink and uplink scenarios. However these linear detectors require matrix inversion which is computationally expensive for such huge systems. In this paper, we assert that calculating an exact inverse is not necessary to find the ZF/MMSE solution and an approximate inverse would yield a similar performance. This is possible if the quantized solution calculated using the approximate inverse is same as the one calculated using the exact inverse. We quantify the amount of approximation that can be tolerated for this to happen. Motivated by this, we propose to employ existing iterative methods for obtaining low complexity approximate inverses. We show that, after a sufficient number of iterations, the inverse using iterative methods can provide a similar error performance. In addition, we also show that the advantage of using an approximate inverse is not limited to linear detectors but can be extended to non linear detectors such as sphere decoders (SD). An approximate inverse can be used for any SD that requires matrix inversion. We prove that application of approximate inverse leads to a smaller radius, which in turn reduces the search space leading to reduction in complexity. Numerical results corroborate our claim that using approximate matrix inversion reduces decoding complexity in large/massive MIMO systems with no loss in error performance.**

I. Introduction

With growing demand for high throughput, Mutiple-Input-Multiple-Output (MIMO) systems with large/massive number of antennas are expected to become an indispensable part of fifth generation wireless technology [1], [2]. It employs a large number of antennas at the base station (of the order of hundreds) that operate to serve relatively fewer users. However, we know that as the number of antennas grow, the complexity of detection algorithms increases [3]. Thus, there is need for techniques which, while exploiting the extra degrees of freedom, are able to decode the transmitted signal efficiently in terms of error performance and complexity.

In the literature, Zero Forcing (ZF) and Minimum Mean Square Error (MMSE) have commonly been used as precoders in a massive MIMO downlink [4], [5] and as decoders in a massive MIMO uplink. Even the complex decoders for uplink transmission also require the computation of ZF/MMSE solution. For example, neighborhood search based algorithms [6], [7] or sparsity based detectors [8], [9] use such linear detectors for initialization. Calculating a ZF or an MMSE solution requires inversion of a matrix. However, finding an inverse is computationally expensive, especially when large number of antennas are employed.

In this paper, we argue that an approximate matrix inverse suffices for finding a ZF/MMSE solution. In other words, usage of an approximate inverse does not compromise the quality of a ZF/MMSE solution. Since the solution obtained using linear detectors anyway needs to be quantized, it is clear that there is a scope for using an approximate inverse as long as the quantized solution remains unchanged. We derive bounds on the approximation such that their quantized ZF/MMSE solutions from the exact and approximate inverses are same in an expected sense. Further, we show that the advantages of using an approximate inversion are not limited to linear detectors. Thus, a class of Sphere Decoding (SD) algorithms [10] require the ZF solution for computing the Babai Radius (BR) [11], [12], consequently requiring matrix inversion. Hence, one can think of utilizing an approximate matrix inverse even in complex decoding schemes like SD.

In this work, we propose the application of an approximate inverse to compute the BR for usage in SD. The approximate inverse has two advantages. Firstly, it reduces the complexity of matrix inversion. But secondly, and more importantly, we prove that it results in a smaller BR. This is a bigger advantage as complexity of decoding in such SD algorithms is largely governed by the choice of BR. Simulations results for large/massive MIMO systems corroborate that the proposed SD provides a low complexity solution with no loss in error performance.

II. System Model

Consider a massive MIMO downlink with N transmit antennas at the base station and K users, each with a single receive antenna. Such a system can be represented by

$$\mathbf{y}_d = \mathbf{H}_d \mathbf{s}_d + \mathbf{n}_d, \tag{1}$$

where $\mathbf{s}_d = \mathbf{W}\mathbf{x}_d$, $\mathbf{W}$ is the linear precoder such as ZF or MMSE and $\mathbf{x}_d$ is the N dimensional signal vector transmitted from the base station. Each element in $\mathbf{x}_d$ is drawn from a set Ω, all entries of which belong to an M-QAM constellation, with average symbol energy E_s. $\mathbf{H}_d$ represents the $K \times N$ channel matrix whose elements are independent and identically distributed (i.i.d.) with zero mean and unit variance, and $\mathbf{n}_d$ is an i.i.d. zero mean Gaussian noise vector with dimension

$K \times 1$ and variance N_0. The i-th entry of the vector $\mathbf{y}_d$, $y_{i,d}$, is the signal intended for the i-th user, for $i = 1, 2, \cdots, K$.

Similarly, in the case of uplink, the system can be represented by

$$\mathbf{y}_u = \mathbf{H}_u \mathbf{x}_u + \mathbf{n}_u, \tag{2}$$

where $\mathbf{x}_u$ is the K dimensional transmitted signal vector whose i-th entry is the symbol transmitted by the i-th user, for $i = 1, 2, \cdots, K$. Again, each element in $\mathbf{x}_u$ is drawn from the set Ω, with average symbol energy E_s. Similarly, $\mathbf{H}_u$ is the $N \times K$ i.i.d. channel matrix with each coefficient having zero mean and unit variance. The noise vector $\mathbf{n}_u$ is i.i.d. $N \times 1$ Gaussian with each element having zero mean and variance N_0, and $\mathbf{y}_u$ is the N dimensional received signal vector at the base station. This results in KE_s/N_0 Signal-to-Noise Ratio (SNR) at each receive antenna.

III. A Linear Detector using Approximate Matrix Inverse

Linear detectors such as ZF and MMSE are useful for both the uplink and downlink (as a precoder) in massive MIMO systems. The expressions for these detectors can be expressed as

$$\mathbf{x}_{\text{ZF}} = \lceil (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \mathbf{y} \rfloor \tag{3}$$

$$\mathbf{x}_{\text{MMSE}} = \left\lceil (\mathbf{H}^H \mathbf{H} + \frac{N_0}{E_s} \mathbf{I}_K)^{-1} \mathbf{H}^H \mathbf{y} \right\rfloor, \tag{4}$$

where $\lceil \cdot \rfloor$ quantization operator to the set Ω and $\mathbf{H}$ is the $N \times K$ channel matrix. Quantization allows us to use an approximate inverse instead of exact inverse while giving the same ZF/MMSE solution. Since the operations are similar in both the uplink and downlink scenarios, we consider only the uplink scenario for the analysis. For notational simplicity, we have removed the subscripts here onwards.

Let us define the error in the approximation of the inverse of a matrix $\mathbf{C}$ as $\mathbf{E} = \widetilde{\mathbf{C}} - \mathbf{C}^{-1}$, where $\mathbf{C} = \mathbf{H}^H \mathbf{H}$ is a $K \times K$ matrix that needs to be inverted and $\widetilde{\mathbf{C}}$ is its approximate inverse. Also, define $\mathbf{g} = \mathbf{H}^H \mathbf{y}$.

A. A Bound on the Acceptable Error in the Matrix Inverse

We will consider an approximate matrix inverse good if the ZF solution calculated through it is equal to that calculated through the exact inverse. In this section, we evaluate a bound on the error which can be tolerated in the computation of an approximate matrix inverse.

For the ZF solutions calculated using the exact and approximate matrix inverses to be equal, the following equality must be satisfied

$$\begin{aligned} \underset{\mathbf{x}\in\Omega}{\text{argmin}} \|\mathbf{x} - \mathbf{C}^{-1}\mathbf{H}^H\mathbf{y}\|^2 &= \underset{\mathbf{x}\in\Omega}{\text{argmin}} \|\mathbf{x} - \widetilde{\mathbf{C}}\mathbf{H}^H\mathbf{y}\|^2 \\ \Rightarrow \underset{\mathbf{x}\in\Omega}{\text{argmin}} \|\mathbf{x} - \mathbf{C}^{-1}\mathbf{g}\|^2 &= \underset{\mathbf{x}\in\Omega}{\text{argmin}} \|\mathbf{x} - \widetilde{\mathbf{C}}\mathbf{g}\|^2 \\ \Rightarrow \underset{\mathbf{x}\in\Omega}{\text{argmin}} \|\mathbf{x} - \mathbf{C}^{-1}\mathbf{g}\|^2 &= \underset{\mathbf{x}\in\Omega}{\text{argmin}} \|\mathbf{x} - \mathbf{C}^{-1}\mathbf{g} - \mathbf{E}\mathbf{g}\|^2. \end{aligned} \tag{5}$$

Let the solution of the L.H.S. of (5) be $\mathbf{x}_{\text{ZF}}$ and let $\mathbf{z} = \mathbf{x}_{\text{ZF}} - \mathbf{C}^{-1}\mathbf{g}$. Therefore, (5) will be satisfied if the following inequalities are satisfied by the error matrix $\mathbf{E}$ (a sufficient condition)

$$-\frac{d_{\min}}{2} < \Re(z_i - \sum_{j=1}^{K} E_{ij} g_j) < \frac{d_{\min}}{2} \tag{6}$$

$$-\frac{d_{\min}}{2} < \Im(z_i - \sum_{j=1}^{K} E_{ij} g_j) < \frac{d_{\min}}{2}, \tag{7}$$

$\forall\ i = 1, 2, \cdots, k$ and $j = 1, 2, \cdots, k$, where z_i is the i-th element of $\mathbf{z}$, E_{ij} is the (i, j)-th element of matrix $\mathbf{E}$, d_{min} is the smallest distance between any two points in the constellation, and $\Re$ and $\Im$ denote the real and imaginary parts respectively. After combining the K equations in (6) and (7) and taking expectations on all sides, we have

$$\frac{-d_{\min}}{2}\mathbf{1}_K < \mathbb{E}[\Re(\mathbf{z} - \mathbf{E}\mathbf{g})] < \frac{d_{\min}}{2}\mathbf{1}_K,$$

$$\frac{-d_{\min}}{2}\mathbf{1}_K < \mathbb{E}[\Im(\mathbf{z} - \mathbf{E}\mathbf{g})] < \frac{d_{\min}}{2}\mathbf{1}_K.$$

where $\mathbf{1}_K$ is a $K \times 1$ vector with all entries as ones.

Here $\mathbb{E}(z_i) = 0$, for all $i = 1, 2, \cdots, N$, because for a given transmitted vector $\mathbf{x}$, $\mathbf{x}_{\text{ZF}}$ can take any point in the constellation around $\mathbf{x}$ due to randomly and independently distributed noise and hence the expectation of difference between the two quantities would be zero. Therefore

$$\frac{-d_{\min}}{2}\mathbf{1}_K < \mathbb{E}[\Re(\mathbf{E}\mathbf{g})] < \frac{d_{\min}}{2}\mathbf{1}_K$$

which, after substituting for $\mathbf{E}$ and $\mathbf{g}$, yields

$$\frac{-d_{\min}}{2}\mathbf{1}_K < \mathbb{E}[\Re((\widetilde{\mathbf{C}}_k - \mathbf{C}^{-1})(\mathbf{H}^H\mathbf{H}\mathbf{x} + \mathbf{H}^H\mathbf{n}))] < \frac{d_{\min}}{2}\mathbf{1}_K$$

$$\Rightarrow \frac{-d_{\min}}{2}\mathbf{1}_K < \mathbb{E}[\Re(\mathbf{S}\mathbf{x})] < \frac{d_{\min}}{2}\mathbf{1}_K, \tag{8}$$

where we define $\mathbf{S} = \mathbf{I} - \widetilde{\mathbf{C}}\mathbf{C}$ as the residual matrix. Similarly,

$$\frac{-d_{\min}}{2}\mathbf{1}_K < \mathbb{E}[\Im(\mathbf{S}\mathbf{x})] < \frac{d_{\min}}{2}\mathbf{1}_K \tag{9}$$

Hence, if (8) and (9) are together satisfied by $\mathbf{S}$ for a given transmitted vector $\mathbf{x}$, the average difference between the ZF solutions through approximate and exact inverses is bounded. This means on an average sense, the two ZF solutions would be same. Next, we discuss some low complexity approximate matrix inversion methods which can be used to find ZF solution accurately.

B. Low Complexity Iterative Methods for Computing Approximate Matrix Inverses

Several low complexity iterative methods for finding the inverse of a matrix have been proposed in [13], [14]. Let $\mathbf{C}_k$ be the approximate inverse and $\mathbf{S}_k$ be the residual matrix after k iterations. The order of the iterative method is p if the residuals after k and $k + 1$ iterations satisfy $\mathbf{S}_{k+1} = \mathbf{S}_k^p$. For e.g., in

a third order method, approximate matrix is calculated in the following manner [13]

$$\mathbf{C}_{k+1} = \mathbf{C}_k(3\mathbf{I} - \mathbf{C}\mathbf{C}_k(3\mathbf{I} - \mathbf{C}\mathbf{C}_k)), \tag{10}$$

where $\mathbf{I}$ is the identity matrix. Here, we note that $\mathbf{S}_{k+1} = \mathbf{S}_k^3$. Similarly, a seventh order iterative method is defined as

$$\begin{aligned}\mathbf{C}_{k+1} &= \mathbf{C}_k(7\mathbf{I} + \mathbf{C}\mathbf{C}_k(21\mathbf{I} + \mathbf{C}\mathbf{C}_k(35\mathbf{I} + \mathbf{C}\mathbf{C}_k(35\mathbf{I} \\ &+ \mathbf{C}\mathbf{C}_k(21\mathbf{I} + \mathbf{C}\mathbf{C}_k(7\mathbf{I} + \mathbf{C}\mathbf{C}_k))))))),\end{aligned} \tag{11}$$

and here, we have $\mathbf{S}_{k+1} = \mathbf{S}_k^7$.

In our simulations, we use Newton's iterative method for finding approximate matrix inverse which has low latency, low complexity [10] and is also easy to implement [14]. The approximate inverse is updated in each iteration according to

$$\mathbf{C}_{k+1} = (2\mathbf{I} - \mathbf{C}_k\mathbf{C})\mathbf{C}_k. \tag{12}$$

Here, $\mathbf{S}_{k+1} = \mathbf{S}_k^2$, revealing quadratic convergence. Increasing the number of iterations increases accuracy, but also increases the number of operations required and hence affects complexity, resulting in a trade-off between performance and efficiency.

Initial matrix $\mathbf{C}_0$ needs to be chosen with care as it decides the number of iterations required for the method to converge, if it converges at all. The applicability of iterative methods is restricted since global convergence is not inherent to all initial matrices. A general condition for initialization is given by $||\mathbf{I} - \mathbf{C}\mathbf{C}_0||_2 < 1$ or $||\mathbf{S}_0||_2 < 1$. This condition ensures that the residual converges towards zero after each iteration.

However, there are some conventional initialization methods which guarantee convergence. In [14], theorem 2 shows that to find the inverse of a matrix $\mathbf{C}$, the initialization $\mathbf{C}_0 = a\mathbf{C}^H$, where a satisfies $0 < a < \frac{2}{\sigma^2_{max}}$ and σ^2_{max} is denoted as the largest eigenvalue of the matrix $\mathbf{A} = \mathbf{C}^H\mathbf{C}$, ensures convergence. To reduce the complexity, following bound is used [14]

$$\sigma^2_{max} \leq \lambda_{upper} = m + t(N-1)^{\frac{1}{2}} \tag{13}$$

where $m = \frac{trace(\mathbf{A})}{N}$ and $t^2 = \frac{trace(\mathbf{A}^2)}{N} - m^2$ and a is selected as $a = 2/\lambda_{upper}$, which ensures convergence. In the next section, we propose a low complexity SD algorithm for large-antenna and massive MIMO systems that uses above matrix inversion methods to accurately estimate the transmitted signal vector.

IV. Sphere Decoding using Iterative Matrix Inverse

Now, let us investigate the advantages of using iterative matrix inverses for non-linear detectors, such as SD. Presently, there are two main versions of SD. The first is the Schnorr-Euchner enumeration [15], [16] that updates the radius for SD adaptively, where after starting with an infinite radius, the search space shrinks with each good point until we get the optimal solution. In large/massive MIMO systems, such a technique would result in a huge decoding complexity. The other one is Fincke-Pohst algorithm based SD [11], [17], which uses a fixed radius approach, and all the points that are inside the search space defined by the radius are compared for detecting the transmitted signal. This technique is extremely sensitive to the choice of the radius. It has been shown in the literature that both these approaches provide near ML performance. In this section, we propose a mechanism to reduce the complexity of SD.

Algorithm 1: Proposed SD Scheme

Input : $\mathbf{y}$, $\mathbf{H}$, Ω, k
Output : $\hat{\mathbf{x}}$

Initialization $i = K$, $cost = r_k$, $\widetilde{c}_i = 0$;
$[\mathbf{Q}\ \mathbf{R}] \leftarrow$ QR decomposition of $\mathbf{H}$ and $\mathbf{z} = \mathbf{Q}^H\mathbf{y}$;
$\hat{\mathbf{x}} \leftarrow \mathbf{DFTS}(\mathbf{z}, \mathbf{R}, \Omega, cost, \widetilde{c}_i, d, i)$;

Function: $\mathbf{DFTS}(\mathbf{z}, \mathbf{R}, \Omega, cost, \widetilde{c}_i, i)$
for $j \leftarrow 1$ **to** *length*(Ω) **do**
  $c_j = |z_i - r_{i,i}x_j|^2$, $\forall x_j \in \Omega$;
end
Sort c_j's in ascending order and keep only those symbols for which $c_j < (cost - \widetilde{c}_i)$;
if $c_j \nleq (cost - \widetilde{c}_i)$ **then**
  return $\hat{\mathbf{x}}$, $cost$;
else
  for $u \leftarrow 1$ **to** *length*$(\mathbf{c})$ **do**
    $\hat{x}_i = x_u$;
    $\widetilde{c}_i \leftarrow \widetilde{c}_i + c_u$;
    if $i = 1$ **then**
      if $cost_{temp} < cost$ **then**
        $cost \leftarrow \widetilde{c}_i$;
        return $\hat{\mathbf{x}}$, $cost$;
      end
    else
      $\widetilde{\mathbf{z}} = \mathbf{z} - \mathbf{R}_{:,u}\,x_u$;
      Extend the tree $\mathcal{T}$ for all Ω;
      $[\hat{\mathbf{x}}, cost] \leftarrow \mathbf{DFTS}(\widetilde{\mathbf{z}}, \mathbf{R}, \Omega, cost, \widetilde{c}_i, i-1)$;
    end
  end
end

Our SD algorithm combines both the strategies wherein we initialize with a BR computed using a low complexity iterative matrix inverse and also update the radius adaptively with every good point. The number of updates when using this algorithm would be significantly less, as the radius will be updated only when a new point is closer to the transmitted signal than ZF. Also, we are always guaranteed a solution as the ZF solution is always inside the searched domain. In Algorithm 1, we show the steps of the proposed SD scheme.

A. Comparison of Babai Radii Calculated through Approximate and Exact Matrix Inverses

Though iterative methods provide a good approximate inverse, it is important to analyze the effect of approximation on the BR, as the choice of radius largely governs the complexity of SD. Interestingly, we show that the application

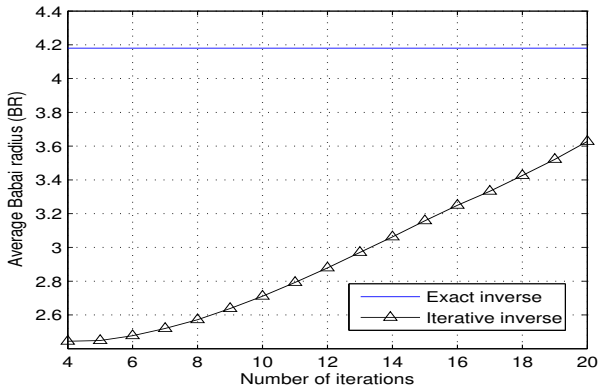

Fig. 1. BR with Newton's iterative method for a 16×16 system.

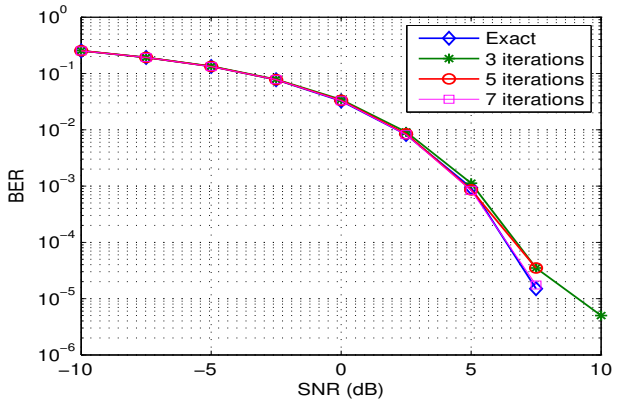

Fig. 2. Bit error performance for the MMSE decoder in a massive MIMO system with $N = 128$, $K = 8$ for 16-QAM.

of approximate matrix inversion methods also reduces the value of radius which leads to further savings in complexity.

To prove this, let us define r_e as the BR computed through exact inverse and r_k as the BR computed through the iterative method after k iterations. We know that $r_e = \lim_{k\to\infty} r_k$. Now, from the definition of BR [11], [12], we can write

$$r_e = ||\mathbf{R}(\mathbf{x}_{\text{ZF}} - \hat{\mathbf{x}}_1)||, \quad (14)$$
$$r_k = ||\mathbf{R}(\mathbf{x}_{\text{ZF}} - \hat{\mathbf{x}}_2)||, \quad (15)$$

where $\hat{\mathbf{x}}_1 = \mathbf{C}^{-1}\mathbf{g}$, $\hat{\mathbf{x}}_2 = \mathbf{C}_k\mathbf{g}$ and $\mathbf{R}$ is obtained from the QR decomposition of $\mathbf{H}$ as $\mathbf{H} = \mathbf{QR}$. Let $\bar{\mathbf{n}}$ denote the noise with respect to $\mathbf{x}_{\text{ZF}}$, i.e.,

$$\mathbf{y} = \mathbf{H}\mathbf{x}_{\text{ZF}} + \bar{\mathbf{n}}. \quad (16)$$

Now, let us define the relation between transmitted vector $\mathbf{x}$ and detected ZF vector $\mathbf{x}_{\text{ZF}}$ as

$$\mathbf{x} = \mathbf{x}_{\text{ZF}} + \Delta, \quad (17)$$

where $||\Delta||$ denotes the magnitude of error in $\mathbf{x}_{\text{ZF}}$. Since $\mathbf{x}$ and $\mathbf{x}_{\text{ZF}}$ both belong to the same constellation, expectation of the difference between $\mathbf{x}$ and $\mathbf{x}_{\text{ZF}}$ would be zero. Substituting (17) in (16), we get $\bar{\mathbf{n}} = \mathbf{n} - \mathbf{H}\Delta$, and thus $\mathbb{E}(\bar{\mathbf{n}}) = 0$.

For a sufficient number of iterations, we can write the expected difference between the squares of the two radii in (14) and (15) as

$$\mathbb{E}[r_e^2 - r_k^2] = 2\Re[\mathbb{E}\{\bar{\mathbf{n}}^H\mathbf{H}\mathbf{S}_k\mathbf{C}^{-1}\mathbf{g}\}]. \quad (18)$$

We prove the above equation in Appendix I.

We next show that the L.H.S. in (18) decreases as the number of iterations increase. Using $\mathbf{g} = \mathbf{H}^H\mathbf{y}$ in (18), we can write

$$\begin{aligned}\mathbb{E}[\bar{\mathbf{n}}^H\mathbf{H}\mathbf{S}_k\mathbf{C}^{-1}\mathbf{g}] &= \mathbb{E}[\bar{\mathbf{n}}^H\mathbf{H}\mathbf{S}_k\mathbf{C}^{-1}\mathbf{H}^H(\mathbf{H}\mathbf{x}_{\text{ZF}} + \bar{\mathbf{n}})] \\ &= \mathbb{E}[\bar{\mathbf{n}}^H\mathbf{H}\mathbf{S}_k\mathbf{x}_{\text{ZF}}] + \mathbb{E}[\bar{\mathbf{n}}^H\mathbf{H}\mathbf{S}_k\mathbf{C}^{-1}\mathbf{H}^H\bar{\mathbf{n}}]. \quad (19)\end{aligned}$$

Also, for a given channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$, $\mathbf{x}_{\text{ZF}}$ would be a constant. Therefore, we can take vector $\mathbf{x}_{\text{ZF}}$ out of the first expectation term in (19) and it becomes

$$\mathbb{E}[\bar{\mathbf{n}}^H\mathbf{H}\mathbf{S}_k\mathbf{x}_{\text{ZF}}] = \mathbb{E}[\bar{\mathbf{n}}^H]\mathbf{H}\mathbf{S}_k\mathbf{x}_{\text{ZF}} = 0, \quad (20)$$

and therefore, can be rewritten as

$$\begin{aligned}\mathbb{E}[\bar{\mathbf{n}}^H\mathbf{H}\mathbf{S}_k\mathbf{C}^{-1}\mathbf{g}] &= \mathbb{E}[\bar{\mathbf{n}}^H\mathbf{H}\mathbf{S}_k\mathbf{C}^{-1}\mathbf{H}^H\bar{\mathbf{n}}] \\ &= N_0\text{Tr}(\mathbf{H}\mathbf{S}_k\mathbf{C}^{-1}\mathbf{H}^H), \quad (21)\end{aligned}$$

where $\text{Tr}(\mathbf{X})$ denotes the trace of matrix $\mathbf{X}$. From (18) and (21)

$$\begin{aligned}\mathbb{E}[r_e^2 - r_k^2] &= 2N_0\Re[\text{Tr}(\mathbf{H}\mathbf{S}_k\mathbf{C}^{-1}\mathbf{H}^H)] \\ &= 2N_0\Re[\text{Tr}(\mathbf{S}_k\mathbf{C}^{-1}\mathbf{H}^H\mathbf{H})] \\ &= 2N_0\Re[\text{Tr}(\mathbf{S}_k)]. \quad (22)\end{aligned}$$

Similarly, for the radius obtained after $k+1$ iterations, we get

$$\mathbb{E}[r_e^2 - r_{k+1}^2] = 2N_0\Re[\text{Tr}(\mathbf{S}_{k+1})]. \quad (23)$$

It can be seen that for the residual matrix $\mathbf{S}_k = \mathbf{I} - \mathbf{C}_k\mathbf{C}$, we have $\text{Tr}(\mathbf{S}_k) \geq 0$. If the iterative methods used for matrix inversion converges to the exact inverse, it can be assumed that $\text{Tr}(\mathbf{S}_k) > \text{Tr}(\mathbf{S}_{k+1})$, as the elements of the residual matrix will tend towards zero as the number of iterations increase. Therefore, from equations (22) and (23), it can be deduced that

$$\begin{aligned}\mathbb{E}[r_e^2 - r_{k+1}^2] &< \mathbb{E}[r_e^2 - r_k^2] \\ \Rightarrow \mathbb{E}[r_{k+1}^2] &> \mathbb{E}[r_k^2]\end{aligned}$$

which means that, in general, BR after k iterations is smaller than the BR calculated after $k+1$ iterations. In Fig. 1, we use Newton's iterative method for computing the approximate inverse and plot the BR for different iterations for a 16×16 MIMO system. A monotonic rise in the value of BR with increasing iterations corroborates the above analysis.

As $r_e = \lim_{k\to\infty} r_k$, therefore

$$\mathbb{E}[r_e^2] > \mathbb{E}[r_k^2] \quad \text{for finite } k,$$

i.e. the BR calculated using the exact inverse is larger than the BR calculated through an iterative method for all the iterations. Thus, as stated before, an approximate inverse can provide twofold savings in complexity.

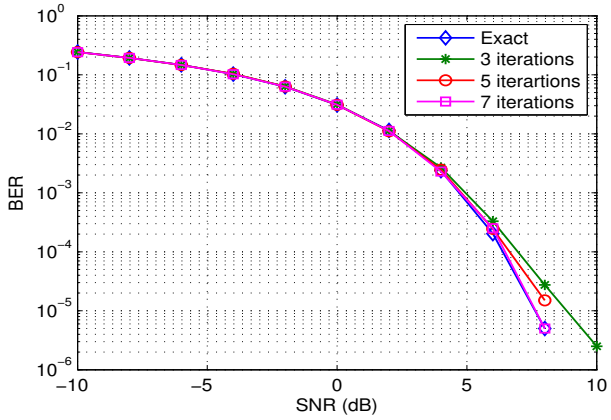

Fig. 3. Bit error performance for the ZF decoder in a massive MIMO system with $N = 128$, $K = 8$ for 16-QAM.

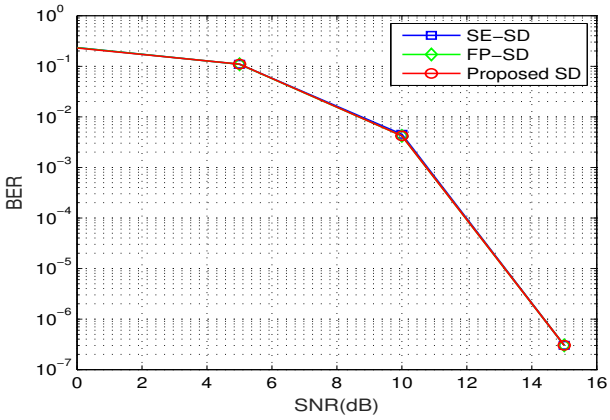

Fig. 5. Bit error performance for different SD schemes for a 16×16 large MIMO system for 4-QAM.

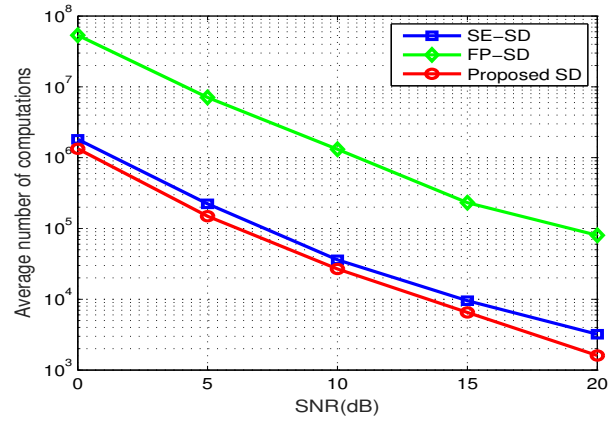

Fig. 4. Average number of computations for different SD schemes for a 16×16 large MIMO system for 4-QAM.

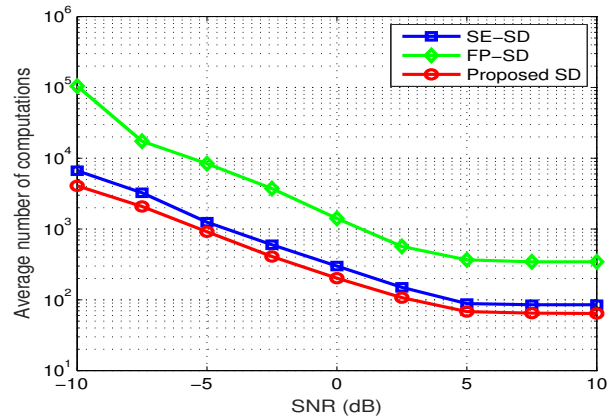

Fig. 6. Average number of computations for different SD schemes for a massive MIMO system with 32 base antennas and 8 users for 4-QAM.

V. Simulation Results

We first examine the performance of ZF and MMSE detectors for massive MIMO scenarios. Subsequently, we compare the simulation results for different SD methods that exist in the literature to the scheme proposed in Algortihm 1. In Fig. 2 and Fig. 3, we plot Bit-Error-Rates (BER) for MMSE and ZF decoders, respectively, for the cases when the matrix inverse is calculated exactly and using Newton's iterative method. We calculate the approximate inverse for 3, 5 and 7 iterations. We see that for 3 and 5 iterations, the error performance in the case of MMSE is slightly away from the case when the exact inverse is used. However, increasing the number of iterations to 7 provides identical performance. Using more number of iterations would not result in any performance gain. Similarly, in the case of ZF decoding, performance improves with the number of iterations, and 7 iterations provides the same performance as the ZF decoder using the exact inverse.

We also perform Monte Carlo simulations for BER and average number of computations for the three different SD schemes discussed above. The first two are adaptive radius (SE-SD) and fixed radius (FP-SD) algorithms respectively. We compare these conventional schemes with the SD scheme proposed in Algorithm 1 in terms of performance and average number of computations required to find the solution. We use Newton's iterative method with 7 iterations to calculate the approximate matrix inverse. In Fig. 4, we compare the average number of computations required by the three schemes for a 16×16 system. It can be observed from the figure that the proposed SD scheme takes at least 35% less number of computations compared to the other two schemes. Also, from Fig. 5, we can deduce that there is no reduction in the quality of performance as all the three schemes give the same BER. In Fig. 6 and Fig. 7, we present similar numerical results for an $N = 32$ and $K = 8$ massive MIMO system. We again note that our SD scheme outperforms the conventional scheme while providing the same error performance.

VI. Conclusion

We have shown the advantages of using an approximate matrix inverse for detectors in large/massive MIMO systems. We obtained the maximum error which can be tolerated in the inverse to arrive at the same quantized ZF/MMSE solution.

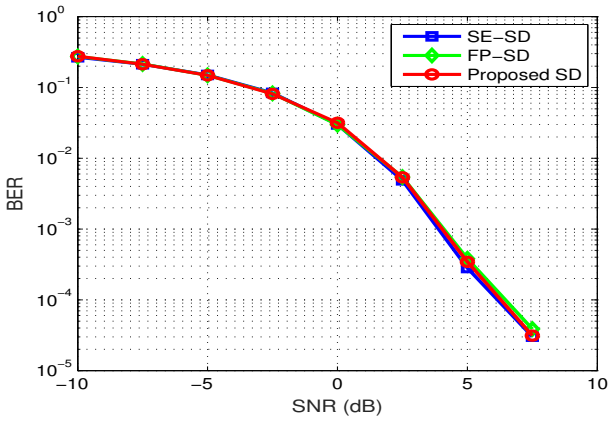

Fig. 7. Bit error performance for different SD schemes for a massive MIMO system with 32 base antennas and 8 users for 4-QAM.

Simulation results show that iterative inversion methods, used to calculate the ZF and MMSE solutions, reached the same performance as provided by the exact inverse for sufficient number of iterations. Extending the idea to complex detectors like SD, we show that the value of BR calculated using iterative methods is less than the BR obtained through the exact method. To this end, we proposed an adaptive SD scheme that uses BR as the initial radius. Simulation results show that the proposed SD scheme outperforms FP-SD and SE-SD in terms of complexity without any loss in performance.

APPENDIX

To prove (18), we use the definition of r_e and r_k from (14) and (15) so that

$$\begin{aligned} r_e^2 - r_k^2 &= ||\mathbf{R}(\hat{\mathbf{x}}_1 - \mathbf{x}_{\text{ZF}})||^2 - ||\mathbf{R}(\hat{\mathbf{x}}_2 - \mathbf{x}_{\text{ZF}})||^2 \\ &= ||\mathbf{R}\mathbf{C}^{-1}\mathbf{g}||^2 - ||\mathbf{R}\mathbf{C}_k\mathbf{g}||^2 \\ &+ 2\{\Re[(\mathbf{R}\mathbf{x}_{\text{ZF}})^H\mathbf{R}(\mathbf{C}_k - \mathbf{C}^{-1})\mathbf{g}]\} \end{aligned}$$

Now, using the fact that $\mathbf{C}_k = \mathbf{C}^{-1} + \mathbf{E}_k$, we get

$$r_e^2 - r_k^2 = 2\Re[(\mathbf{x}_{\text{ZF}} - \mathbf{C}^{-1}\mathbf{g})^H\mathbf{R}^H\mathbf{R}\mathbf{E}_k\mathbf{g}] - ||\mathbf{R}\mathbf{E}_k\mathbf{g}||^2. \quad (24)$$

After using (16) and taking expectations on both sides, we get

$$\mathbb{E}[r_e^2 - r_k^2] = \mathbb{E}[2\Re\{(-\mathbf{C}^{-1}\mathbf{H}^H\bar{\mathbf{n}})^H\mathbf{R}^H\mathbf{R}\mathbf{E}_k\mathbf{g}\} - ||\mathbf{R}\mathbf{E}_k\mathbf{g}||^2]. \quad (25)$$

We will be neglecting the second term in R.H.S. of (25) citing the following assertion

$$||\mathbf{R}\mathbf{E}_k\mathbf{g}||^2 = (\mathbf{R}\mathbf{E}_k\mathbf{g})^H(\mathbf{R}\mathbf{E}_k\mathbf{g}) = \mathbf{g}^H\mathbf{E}_k^H\mathbf{R}^H\mathbf{R}\mathbf{E}_k\mathbf{g}. \quad (26)$$

From the orthogonal property of $\mathbf{Q}$, we have $\mathbf{R}^H\mathbf{R} = \mathbf{H}^H\mathbf{H} = \mathbf{C}$ and therefore (26) becomes

$$||\mathbf{R}\mathbf{E}_k\mathbf{g}||^2 = \mathbf{g}^H\mathbf{E}_k^H\mathbf{C}\mathbf{E}_k\mathbf{g}$$

Using $\mathbf{E}_k = \mathbf{C}_k - \mathbf{C}^{-1}$, we have

$$\begin{aligned} ||\mathbf{R}\mathbf{E}_k\mathbf{g}||^2 &= \mathbf{g}^H(\mathbf{C}_k^H - (\mathbf{C}^{-1})^H)\mathbf{C}(\mathbf{C}_k - \mathbf{C}^{-1})\mathbf{g} \\ &= \mathbf{g}^H(\mathbf{C}_k^H\mathbf{C}^H - \mathbf{I})(\mathbf{C}_k\mathbf{C} - \mathbf{I})\mathbf{C}^{-1}\mathbf{g} \\ &= \mathbf{g}^H\mathbf{S}_k^H\mathbf{S}_k\hat{\mathbf{x}}_1, \end{aligned}$$

where $\mathbf{S}_k = \mathbf{I} - \mathbf{C}_k\mathbf{C}$ is the residual matrix. Here, we have used the fact that $\mathbf{C}$ is a Hermitian matrix and $(\mathbf{C}^{-1})^H = \mathbf{C}^{-1}$. Also, the first term in the R.H.S. of (25) can be written as

$$\begin{aligned} \mathbb{E}[2\Re\{(-\mathbf{C}^{-1}\mathbf{H}^H\bar{\mathbf{n}})^H\mathbf{R}^H\mathbf{R}\mathbf{E}_k\mathbf{g}\}] &= \mathbb{E}[2\Re(\bar{\mathbf{n}}^H\mathbf{H}\mathbf{E}_k\mathbf{g})] \\ &= \mathbb{E}[2\Re(\bar{\mathbf{n}}^H\mathbf{H}\mathbf{S}_k\mathbf{C}^{-1}\mathbf{g})]. \end{aligned}$$

For sufficient number of iterations, $\mathbf{S}_k$ would be very small and hence the term $\mathbb{E}[||\mathbf{R}\mathbf{E}_k\mathbf{g}^2||]$ can be neglected when compared to first term in the R.H.S. of equation (25), as the former is proportional to $\mathbf{S}_k^H\mathbf{S}_k$ while the latter is proportional to $\mathbf{S}_k$. Hence, (25) can be rewritten as

$$\mathbb{E}[r_e^2 - r_k^2] \approx \mathbb{E}[2\Re(\bar{\mathbf{n}}^H\mathbf{H}\mathbf{S}_k\mathbf{C}^{-1}\mathbf{g})]$$

which proves (18) for sufficient number of iterations k.

REFERENCES

[1] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 186–195, February 2014.

[2] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 74–80, February 2014.

[3] L. Lu, G. Li, A. Swindlehurst, A. Ashikhmin, and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *Selected Topics in Signal Proc., IEEE Journal of*, vol. 8, no. 5, pp. 742–758, Oct 2014.

[4] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of Cellular Networks: How Many Antennas Do We Need?" *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 2, pp. 160–171, February 2013.

[5] H. Li and V. Leung, "Low complexity zero-forcing beamforming for distributed massive MIMO systems in large public venues," *Comm. and Networks, Journal of*, vol. 15, no. 4, pp. 370–382, Aug 2013.

[6] A. K. Sah and A. K. Chaturvedi, "Reduced neighborhood search algorithms for low complexity detection in MIMO systems," in *GLOBECOM 2015. IEEE International Symposium on*, Dec 2015.

[7] N. Srinidhi, T. Datta, A. Chockalingam, and B. Rajan, "Layered Tabu Search Algorithm for Large-MIMO Detection and a Lower Bound on ML Performance," *Communications, IEEE Transactions on*, vol. 59, no. 11, pp. 2955–2963, November 2011.

[8] X. Peng, W. Wu, J. Sun, and Y. Liu, "Sparsity-boosted detection for large MIMO systems," *Communications Letters, IEEE*, vol. 19, no. 2, pp. 191–194, Feb 2015.

[9] J. W. Choi and B. Shim, "New approach for massive MIMO detection using sparse error recovery," in *Global Communications Conference (GLOBECOM), 2014 IEEE*, Dec 2014, pp. 3754–3759.

[10] Y. Wang and H. Leib, "Sphere decoding for mimo systems with newton iterative matrix inversion," *Communications Letters, IEEE*, vol. 17, no. 2, pp. 389–392, February 2013.

[11] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. Expected complexity," *Signal Processing, IEEE Transactions on*, vol. 53, no. 8, pp. 2806–2818, Aug 2005.

[12] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *Information Theory, IEEE Transactions on*, vol. 48, no. 8, pp. 2201–2214, Aug 2002.

[13] F. Soleymani, "On a Fast Iterative method for Approximate Inverse of Matrices," *Communications of the Korean Mathematical Society*, 2013.

[14] V. Pan and R. Schreiber, "An improved newton iteration for the generalized inverse of a matrix, with applications," *SIAM Journal on Scientific and Statistical Computing*, vol. 12, no. 5, pp. 1109–1130, 1991.

[15] T. Cui, S. Han, and C. Tellambura, "Probability-Distribution-Based Node Pruning for Sphere Decoding," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 4, pp. 1586–1596, May 2013.

[16] C. Schnorr and M. Euchner, "Lattice basis reduction: Improved practical algorithms and solving subset sum problems," *Mathematical Programming*, vol. 66, no. 1-3, pp. 181–199, 1994.

[17] M. P. U. Fincke, "Improve Methods for Calculating Vectors of Short Length in a lattice, including a complexity analysis," *Math. Comp.*, 1985.