Low Complexity Iterative MMSE-PIC Detection for Medium-Size Massive MIMO

Licai Fang, *Student Member, IEEE,* Lu Xu, *Member, IEEE,* and Defeng (David) Huang, *Senior Member, IEEE*

***Abstract*—In medium-size Massive MIMO systems, the minimum mean square error parallel interference cancellation (MMSE-PIC) based Soft-Input Soft-Output (SISO) detector is often used due to its relatively low complexity and good bit error rate (BER) performance. The computational complexity of MMSE-PIC for detecting a block of data is dominated by the computation of a Gram matrix and a matrix inversion. They have computational complexity of $\mathcal{O}(K^2M)$ and $\mathcal{O}(K^3)$, respectively, where K is the number of uplink users with one transmit antenna each and M is the number of receive antennas at the base station. In this letter, by using an L (typically $L \leq 3$) terms of Neumann series expansion to approximate the matrix inversion, we reduce the *total* computational complexity to $\mathcal{O}(LKM)$. Compared with alternative algorithms which focus on reducing the complexity of the matrix inversion only, the proposed method can also avoid calculating the Gram matrix explicitly and thus significantly reducing the *total* complexity.**

***Index Terms*—Low complexity, Massive MIMO, Neumann series expansion, iterative detection, MMSE.**

I. Introduction

In recent years, Massive MIMO which typically employs a magnitude of more antennas at the base station than in user terminals has attracted great interest from wireless communication research community [1]. It has been shown that with Massive MIMO, the throughput and spectral efficiency of wireless systems can be greatly improved [2]. When the number of receive antennas at the base station is large and much larger than the number of total transmit antennas in user terminals, a simple detection algorithm such as a matched filter can achieve very good performance, as with the assumption of i.i.d. entries for channel matrix $\mathbf{H}$, the channel vectors become orthogonal to each other and $\mathbf{H}^{\mathrm{H}}\mathbf{H}$ converges to a scaled identity matrix. But for practical medium-size Massive MIMO, matched filter based detection algorithm suffers performance loss [3]. Therefore, alternative linear detection algorithms such as the minimum mean square error parallel interference cancellation (MMSE-PIC) algorithm [4] are often employed due to their relatively low complexity and good bit error rate (BER) performance. However, the MMSE-PIC still requires complexity of $\mathcal{O}(K^3)$ for calculating a matrix inversion and $\mathcal{O}(K^2M)$ for calculating the Gram matrix, where K is the number of transmit antennas and M is the number of receive antennas.

To reduce the complexity, [5] and [6] employed Neumann series expansion to approximate the matrix inversion by a matrix polynomial. Then in [3] the authors proposed to use the same method to perform 3GPP-LTE uplink signal detection and proved the convergence of the Neumann series expansion. Different from using Neumann series expansion, in [7] an iterative method based on successive overrelaxation (SOR) is employed to calculate the product of the inversion of a matrix and a vector, which can converge to the exact solution. These work can successfully reduce the complexity of computing matrix inversion from $\mathcal{O}(K^3)$ to $\mathcal{O}(K^2)$. But they all require the pre-computed Gram matrix as an input. In Massive MIMO with $M \gg K$, the Gram matrix computation involves computational complexity of $\mathcal{O}(K^2M)$, which is much higher than the $\mathcal{O}(K^3)$ complexity of matrix inversion .

In this letter, based on the MMSE detection algorithm [8], we exploit Neumann series expansion to reduce the *total* complexity of MMSE-PIC for Massive MIMO. With the proposed method, computational complexity is reduced by avoiding direct matrix inversion and replacing the matrix-matrix multiplication of Gram matrix with matrix-vector multiplications. Specifically, we propose to employ an L (typically $L \leq 3$) terms Neumann series expansion for calculating the means of data symbols to be detected, and a first order approximation for calculating the variances and thus reducing the complexity from $\mathcal{O}(K^2M+K^3)$ to $\mathcal{O}(LKM)$ with marginal performance loss when $L = 3$ for MIMO size of $K \times M = 16 \times 128$. We also investigate the application of the proposed algorithm in an iterative detection and decoding (IDD) system, where the symbol detector and the channel decoder work iteratively. We found that with one iteration between the decoder and the detector, the proposed approximation algorithm with $L = 3$ can achieve the same performance as the exact MMSE-PIC algorithm.

The remainder of this letter is organized as follows. Section II describes the turbo-MIMO system model. Then in Section III, we propose to use Neumann series expansion to perform MMSE detection without computing the Gram matrix. Simulation results are shown in Section IV and Section V concludes this letter.

The notations used in this letter are as follows. Lower and upper case letters denote scalars. Bold lower and upper case letters represent column vectors and matrices, respectively.

The superscripts "T" and "H" denote the transpose and conjugate transpose, respectively.

II. System Model

Consider a multiuser Massive MIMO system with M receive antennas at the base station and K single-antenna user terminals. Let $\mathbf{x} = [x_1, x_2, \ldots, x_K]^{\mathrm{T}}$ denote the trans-

This work was supported by Australian Research Councils Discovery Project DP140100522 and also by iVEC (http://www.ivec.org) through the use of advanced computing resources.

mit vector comprising the symbols transmitted simultaneously by all users in one channel use where $x_n \in \mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_{2^Q}\}(|\mathcal{A}| = 2^Q)$ denotes transmitted symbol from user n, then each x_n corresponds to a length-Q subsequence of $\mathbf{c}$ denoted by $\mathbf{c}_n = [c_{n,1}, c_{n,2}, \cdots, c_{n,Q}]^T$. Let $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K]$ denote the channel gain matrix, where $\mathbf{h}_n = [h_{1n}, h_{2n}, \ldots, h_{Mn}]^T$ is the channel gain vector from user n to the base station, and h_{jn} denotes the channel gain from the n-th user to the j-th receive antenna at the base station. Assuming rich scattering, adequate spatial separation between the base station antenna elements and perfect user power control, $h_{jn}, \forall j$ are assumed to be i.i.d. complex Gaussian distributed with zero mean and variance one. Thus a length-M observation vector $\mathbf{y}$ at the base station can be written as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w} \tag{1}$$

where $\mathbf{w}$ denotes a length-M circularly symmetric additive white Gaussian noise (AWGN) vector with zero-mean and covariance of $\sigma^2\mathbf{I}$.

The task of the Soft-In Soft-Out (SISO) detector is to compute the extrinsic log-likelihood ratio (LLR) for each code bit $c_{n,q}$, which is the input to the decoder and can be expressed as [8]

$$L^{\mathrm{e}}(c_{n,q}) = \ln \frac{\sum\limits_{x_n \in \mathcal{A}_q^0} P(\mathbf{y}|x_n)P(x_n)}{\sum\limits_{x_n \in \mathcal{A}_q^1} P(\mathbf{y}|x_n)P(x_n)} - L^{\mathrm{a}}(c_{n,q}) \tag{2}$$

where $L^{\mathrm{a}}(c_{n,q})$ is the output extrinsic LLR of the decoder, $x_n \in \mathcal{A}_q^0(\mathcal{A}_q^1)$ represents constellations whose q-th bit is $0(1)$ and $P(x_n)$ is the *a priori* probability of x_n which can be calculated from $L^{\mathrm{a}}(c_{n,q})$.

III. MMSE Detection Based on Neumann Series Expansion

We employ the method proposed in [8] to perform MIMO MMSE detection. With this algorithm, it is easy to reformulate the matrix to be inverted with the size of $K \times K$ which is preferable for Massive MIMO applications with $M \gg K$. The core part of this algorithm is to compute the *a posteriori* mean $\mathbf{m}^{\mathrm{p}}$ and variance $\mathbf{V}^{\mathrm{p}}$ of $\mathbf{x}$ by

$$\mathbf{V}^{\mathrm{p}} = (\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{H}^{\mathrm{H}}\mathbf{H})^{-1}, \tag{3}$$

$$\mathbf{m}^{\mathrm{p}} = \mathbf{m} + \frac{1}{\sigma^2}\mathbf{V}^{\mathrm{p}}(\mathbf{H}^{\mathrm{H}}\mathbf{y} - \mathbf{H}^{\mathrm{H}}\mathbf{H}\mathbf{m}), \tag{4}$$

where $\mathbf{m}$ and $\mathbf{V}$ are the *a priori* mean and variance of $\mathbf{x}$, respectively, and they can be calculated from the feedback of the decoder[1]. Then the *extrinsic* mean m_n^{e} and variance v_n^{e} of the n-th element of $\mathbf{x}$ (which are used to generate soft-out LLR) can be calculated by

[1]At the beginning of the IDD, there is no feedback from the decoder. Assuming that the constellation of the modulation is with zero mean and normalized with unit power and data streams from different transmit antennas are statistically independent, we have $\mathbf{m}$ be a zero vector and $\mathbf{V}$ be the identity matrix $\mathbf{I}_K$ with size $K \times K$.

Algorithm 1 Reduced Complexity Neumann Series expansion based MMSE detection

Input: $\mathbf{y}$, $\mathbf{H}$, $\mathbf{L}^{\mathrm{a}}$
Output: $\mathbf{L}^{\mathrm{e}}$ ▷ *extrinsic* LLR value for every bit

Calculate *a priori* mean $\mathbf{m}$ and variance $\mathbf{V}$ from $\mathbf{L}^{\mathrm{a}}$
  $m_n = \sum\limits_{\alpha_i \in \mathcal{A}} \alpha_i P(x_n = \alpha_i)$
  $v_n = \sum\limits_{\alpha_i \in \mathcal{A}} |\alpha_i - m_n|^2 P(x_n = \alpha_i)$
Calculate *a posteriori* mean $\mathbf{m}^{\mathrm{p}}$
  $\mathbf{D} = diag(\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{H}^{\mathrm{H}}\mathbf{H})$
  $\mathbf{v}_0 = \mathbf{D}^{-1}(\mathbf{H}^{\mathrm{H}}\mathbf{y} - \mathbf{H}^{\mathrm{H}}\mathbf{H}\mathbf{m})$
  $\mathbf{s}_0 = \mathbf{v}_0$
  for $i = 1$ to L **do**
    $\mathbf{v}_i = \mathbf{v}_{i-1} - \mathbf{D}^{-1}(\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{H}^{\mathrm{H}}\mathbf{H})\mathbf{v}_{i-1}$
    $\mathbf{s}_i = \mathbf{s}_{i-1} + \mathbf{v}_i$
  end for
  $\mathbf{m}^{\mathrm{p}} = \mathbf{m} + \frac{1}{\sigma^2}\mathbf{s}_L$
Approximate the diagonal elements of $\mathbf{V}^{\mathrm{p}}$
  $v_n^{\mathrm{p}} = d_n$ ▷ d_n is the (n,n)-th element of $\mathbf{D}^{-1}$
Calculate *extrinsic* mean m_n^{e} and variance v_n^{e}
  $v_n^{\mathrm{e}} = (\frac{1}{v_n^{\mathrm{p}}} - \frac{1}{v_n})^{-1}$
  $m_n^{\mathrm{e}} = v_n^{\mathrm{e}}(\frac{m_n^{\mathrm{p}}}{v_n^{\mathrm{p}}} - \frac{m_n}{v_n})$
Calculate *extrinsic* LLR $\mathbf{L}^{\mathrm{e}}$
  $L^{\mathrm{e}}(c_{n,q}) = \ln \frac{\sum\limits_{\alpha_i \in \mathcal{A}_q^0} \exp\left(-\frac{|\alpha_i - m_n^{\mathrm{e}}|^2}{v_n^{\mathrm{e}}}\right) \prod\limits_{q' \neq q} P(c_{n,q'} = s_{i,q'})}{\sum\limits_{\alpha_i \in \mathcal{A}_q^1} \exp\left(-\frac{|\alpha_i - m_n^{\mathrm{e}}|^2}{v_n^{\mathrm{e}}}\right) \prod\limits_{q' \neq q} P(c_{n,q'} = s_{i,q'})}$

$$v_n^{\mathrm{e}} = (\frac{1}{v_n^{\mathrm{p}}} - \frac{1}{v_n})^{-1}, \tag{5}$$

$$m_n^{\mathrm{e}} = v_n^{\mathrm{e}}(\frac{m_n^{\mathrm{p}}}{v_n^{\mathrm{p}}} - \frac{m_n}{v_n}), \tag{6}$$

where v_n, v_n^{p} are the (n,n)-th elements of matrix $\mathbf{V}$ and $\mathbf{V}^{\mathrm{p}}$, respectively, and m_n, m_n^{p} are the n-th elements of vector $\mathbf{m}$ and $\mathbf{m}^{\mathrm{p}}$, respectively. It is easy to see that (3) and (4) require a computational complexity of $\mathcal{O}(K^2M)$ for calculating $\mathbf{H}^{\mathrm{H}}\mathbf{H}$ and $\mathcal{O}(K^3)$ for calculating the matrix inverse.

A. Neumann Series Expansion

The convergence of Neumann series expansion for detection has been proved in [3]. It has been shown in [3] that, for large $\rho = M/K$, the Gram matrix $\mathbf{G} = \mathbf{H}^{\mathrm{H}}\mathbf{H}$ tends to be diagonally dominant, which enables the convergence of the Neumann series expansion.

Let us decompose the regularized Gram matrix $\mathbf{A} = \mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{G}$ to $\mathbf{A} = \mathbf{D} + \mathbf{E}$, where $\mathbf{D}$ is the main diagonal of $\mathbf{A}$. As $\mathbf{V}$ is a diagonal matrix, the complexity of computing $\mathbf{D}$ is the same as computing the diagonal elements of $\mathbf{G}$. We can then approximate $\mathbf{A}^{-1}$ in the Neumann series as

$$\begin{aligned}\mathbf{A}^{-1} &\approx \sum_{i=0}^{L} \left(\mathbf{I}_K - \mathbf{D}^{-1}\mathbf{A}\right)^i \mathbf{D}^{-1} \\ &= \sum_{i=0}^{L} \left(\mathbf{I}_K - \mathbf{D}^{-1}\mathbf{V}^{-1} - \frac{1}{\sigma^2}\mathbf{D}^{-1}\mathbf{G}\right)^i \mathbf{D}^{-1}.\end{aligned} \tag{7}$$

Using $\mathbf{A}^{-1}$ of (7) to replace $\mathbf{V}^{\mathrm{P}}$ and plugging it into the representation of $\mathbf{m}^{\mathrm{P}}$ of (4), it can be seen that only matrix-vector multiplications are needed for calculating $\mathbf{m}^{\mathrm{P}}$ and the calculation of the Gram matrix $\mathbf{G}$ itself is avoided. But we should note that in (5) and (6) the diagonal elements of $\mathbf{V}^{\mathrm{P}}$ are also required to compute the *extrinsic* mean and variance. To reduce the complexity, we propose to use the first order approximation ($L = 0$) of (7) for computing the diagonal elements of $\mathbf{V}^{\mathrm{P}}$ (i.e. $\mathbf{V}^{\mathrm{P}} \approx \mathbf{D}^{-1}$).

From (7), it is obvious that the multiplication of $\mathbf{A}^{-1}$ and a vector $\mathbf{v}$ can be computed by L loops. The proposed MIMO MMSE detection algorithm with Neumann series expansion is summarized in **Algorithm 1**. We note that when $L = 0$, the proposed algorithm coincides with the matched filter detector as $\mathbf{m}^{\mathrm{P}} = \frac{1}{\sigma^2}\mathbf{D}^{-1}\mathbf{H}^{\mathrm{H}}\mathbf{y}$ (Note that we assume $\mathbf{m}$ is a zero vector at the beginning of IDD).

B. Computational Complexity Comparison

We focus on the number of real-valued multiplications needed and only count quadratic or beyond terms. For the real-valued system model, the matrix size of $\mathbf{H}$ is $2K \times 2M$, $\mathbf{y}$ is a length-$2M$ vector and $\mathbf{m}$ is a length-$2K$ vector. Note that using the symmetric property of matrix $\mathbf{G}$ and $\mathbf{V}^{\mathrm{P}}$ can reduce the complexity by a half. **Table I** is a summary of complexity comparison between MMSE, the proposed algorithm, Neumann series expansion based algorithm in [5] and SOR based algorithm in [7]. In the table, the term $4K^2M$ corresponds to the computing of Gram matrix $\mathbf{G}$. Note that for SOR based algorithm in [7], the number of iterations L_s may be smaller than that of Neumann series expansion.

TABLE I
COMPUTATIONAL COMPLEXITY COMPARISON

Algorithm	Number of multiplications
Exact MMSE [8]	$8K^2 + 4K^3 + 4(K^2 + K)M$
Proposed	$(16 + 8L)KM$
Neumann series based [5]	$4K^2M + 8(L-2)K^3$
SOR based [7]	$4K^2M + 4L_sK^2$

C. Discussion

In contrast to [5], [6], and [3], which also use the Neumann series expansion to approximate matrix inversion, the proposed methods avoid direct matrix inversion and replace the matrix-matrix multiplication by matrix-vector multiplications, which result in considerable saving in computations.

The method proposed in [7], after optimizing a parameter by off-line exhaustive searching, can converge faster than Neumann series expansion. But it requires each element of matrix $\mathbf{G}$ as its input, which means that $\mathbf{H}^{\mathrm{H}}\mathbf{H}$ has to be computed explicitly, thus it cannot reduce the total complexity significantly.

IV. SIMULATION RESULTS

We consider a Rayleigh block fading random channel where $\mathbf{H}$ does not change over a codeword. During simulations, we assume that perfect channel information is available in the detection module. A rate-1/2, regular (3,6) low-density parity-check (LDPC) code with codeword length of 2000 bits is employed as the channel code and the maximum number of iterations of the decoder is 25. The constellation of 64-QAM with Gray mapping is used. We constrain the total transmitter power to one, and set the noise variance at each receive antenna to σ^2. Then the average received signal-to-noise ratio (SNR) at each receive antenna is given by $1/\sigma^2$. For each SNR value, we simulate at least 100000 codewords. In the simulations, clipping is applied to both the soft-output and the soft-input of the detector. The soft-in clipping threshold[2] for the *a priori* LLR is ± 2, and soft-output module constrains the output LLR range to $[-50, 50]$.

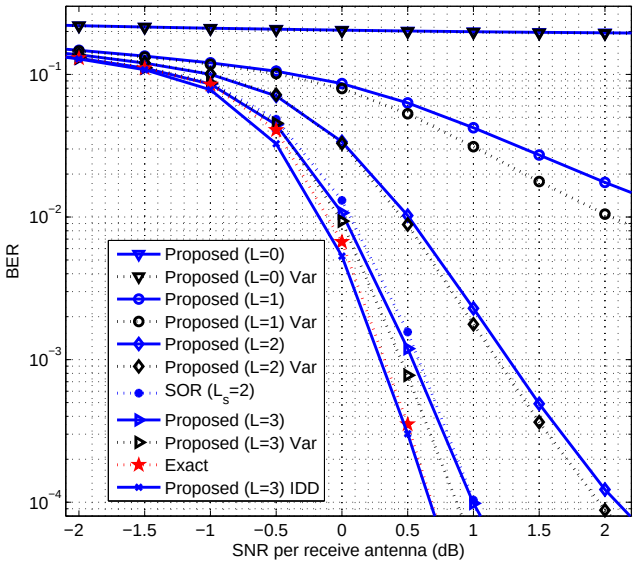

Fig. 1. BER performance comparison for exact MMSE, proposed and SOR based [7] with MIMO size of $K \times M = 16 \times 128$

Fig. 1 shows the BER performance comparison between the exact MMSE detection [8], the proposed algorithm and the SOR based algorithm [7]. The MIMO size is $K \times M = 16 \times 128$. It is easy to see that the performance of the matched filter (with legend *Proposed (L=0)*) is poor. At the same time, with a larger L the approximation is more accurate and when $L = 3$ the proposed algorithm can approach the performance of the exact algorithm within 0.3dB. It can also be seen that an extra IDD iteration (with legend *Proposed (L=3) IDD*) achieves slightly better performance than the exact MMSE-PIC algorithm without IDD.

To evaluate the performance loss caused by the first order approximation of $\mathbf{V}^{\mathrm{P}}$, we use (7) to explicitly compute the matrix inversion and assign the diagonal elements to v_n^{P} (as in [5]) and the performances are shown in Fig. 1 with legends ending with *Var*. It is obvious that the proposed approximation to variance only leads to a small performance penalty.

[2]This clipping threshold can also help resolve the numerical stability issue of Line 16 and Line 17 of **Algorithm 1** when the *a priori* variance v_n is close to zero.

V. Conclusion

In this letter, we have proposed to use Neumann series expansion to reduce the complexity of the MMSE-PIC algorithm for Massive MIMO applications with $M \gg K$. Firstly, an L terms Neumann series was employed to avoid computing the matrix inversion by replacing it with a cascade of matrix-vector multiplications. Then, a first-order approximation was employed to compute the diagonal elements of the *a posteriori* variance matrix for calculating LLR, which helps to avoid computing the Gram matrix explicitly. Simulation results showed that with a small L the proposed approximation methods lead to marginal performance loss compared with the exact implementation, but with considerable complexity saving.

References

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

[2] F. Rusek, D. Persson, B. K. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, 2013.

[3] M. Wu, B. Yin, G. Wang, C. Dick, J. Cavallaro, and C. Studer, "Large-scale MIMO detection for 3GPP LTE: Algorithms and FPGA implementations," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 916–929, 2014.

[4] C. Studer, S. Fateh, and D. Seethaler, "ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation," *IEEE J. Solid-State Circuits*, vol. 46, no. 7, pp. 1754–1765, 2011.

[5] M. Wu, B. Yin, A. Vosoughi, C. Studer, J. Cavallaro, and C. Dick, "Approximate matrix inversion for high-throughput data detection in the large-scale MIMO uplink," in *Proc. IEEE ISCAS,* 2013, pp. 2155–2158.

[6] B. Yin, M. Wu, C. Studer, J. Cavallaro, and C. Dick, "Implementation trade-offs for linear detection in large-scale MIMO systems," in *Proc. IEEE ICASSP*, 2013, pp. 2679–2683.

[7] 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. IEEE GLOBECOM*, 2014, pp. 3291–3295.

[8] Q. Guo and D. Huang, "A concise representation for the soft-in soft-out lmmse detector," *IEEE Commun. Lett.*, vol. 15, no. 5, pp. 566–568, 2011.