Low-Complexity Joint Weighted Neumann Series and Gauss-Seidel Soft-Output Detection for Massive MIMO Systems

Xiaoming Dai, Tiantian Yan, Yuanyuan Dong, Yuquan Luo, and Hua Li

Abstract

We introduce a joint weighted Neumann series (WNS) and Gauss-Seidel (GS) approach to implement an approximated linear minimum mean-squared error (LMMSE) detector for uplink massive multiple-input multiple-output (M-MIMO) systems. We first propose to initialize the GS iteration by a WNS method, which produces a closer-to-LMMSE initial solution than the conventional zero vector and diagonal-matrix based scheme. Then the GS algorithm is applied to implement an approximated LMMSE detection iteratively. Furthermore, based on the WNS, we devise a low-complexity approximate log-likelihood ratios (LLRs) computation method whose performance loss is negligible compared with the exact method. Numerical results illustrate that the proposed joint WNS-GS approach outperforms the conventional method and achieves near-LMMSE performance with significantly lower computational complexity.

Index Terms–Linear minimum mean-squared error (LMMSE), Gauss-Seidel (GS), weighted Neumann series (WNS), massive multiple-input multiple-output (M-MIMO).

I. Introduction

Massive MIMO (M-MIMO) systems can significantly improve the link reliability and spectral efficiency compared to the small-scale MIMO systems. The theoretically predicted gains of the M-MIMO rely on optimal multi-user signal separation at the receiver. In the uplink channel, the optimality of the maximum ratio combining (MRC) is guaranteed theoretically for M-MIMO systems with an infinite number of BS antennas. However, systems with realistic antenna

X. Dai, T. Yan, Y. Dong, Y. Luo and H. Li are with University of Science and Technology Beijing, Beijing 100030 (email: daixiaoming@ustb.edu.cn).

configurations are far from the infinite-antenna limit, e.g, 64 Basestation antenna, which renders the MRC detector not attractive in practical applications. The advent of the Internet of Things (IoT) will foresee a large number of active users in future systems. Therefore it is highly desirable to design low-complexity high-performance detectors for practical “not-so-massive systems” with a low base station (BS)-to-user-antenna ratio (BUAR). The maximum-likelihood (ML) detector can achieve optimal performance, however, its complexity increases exponentially with the number of users, which makes it unaffordable for M-MIMO systems with large number of users. The sphere decoding [2] can achieve near-optimal performance with reduced average complexity. Nevertheless, its worst-case complexity is still prohibitive when the dimension of the MIMO system is large and/or the modulation order is high. Furthermore, the variable complexity of the SD makes it challenging for practical hardware implementation. The linear minimum mean-squared error (LMMSE) detector has been widely utilized in multiple communication applications due to its good performance-complexity tradeoff. The LMMSE detector is shown to achieve near-optimal error-rate performance for M-MIMO systems with a large BUAR [1]. However, the associated matrix-inversion entails high computational complexity for practical implementation with a large number of users. To address the complexity issue, the Richardson method has been proposed in [3], but its performance is highly sensitive to the step-size, thus rendering it less attractive for practical implementation. By exploiting the channel hardening property of the M-MIMO channel, the Neumann series (NS) approximation approach [4] is proposed to convert the matrix inversion into a series of matrix-vector multiplications to alleviate the implementation burden of the matrix inversion. The performance of the NS method approaches that of the LMMSE for a M-MIMO system with a large BUAR. However, the performance of the NS-based approach degrades significantly if the BUAR is not large enough, e.g., less than four [5]. Dai *et al* [6] proposed a Gauss-Seidel (GS) method to implement an approximated LMMSE detection. The GS-based approach [6] exhibits a slow convergence rate, especially for the system with a low BUAR.

In this work, we propose a joint weighted Neumann series (WNS) and Gauss-Seidel approach to implement the LMMSE detection without matrix inversion. In the proposed scheme, a two-term NS is first utilized to generate an initial solution for the subsequent GS-based iterative detection. To further improve the quality of the initial solution, a weighted approach is introduced to compensatefor the approximation error introduced by the limited expansion term of the NS. The

weighting coefficients are determined empirically based on an off-line basis. The GS iterative algorithm then exploits the promising initial solution to achieve a faster convergence rate. We also propose a low-complexity LLR computation based on the WNS with only negligible performance degradation. Numerical results illustrate that the proposed method outperforms the GS method [6] and approaches the performance of the LMMSE detector with significantly lower computational complexity.

The rest of this work is organized as follows. Section II describes the system model.The proposed joint WNS and GS method is detailed In Section III. The performance and complexity of the proposed method is compared with the conventional ones in Section IV. The concluding remarks are provided in Section V. The system model is and the basic principle of the NS and GS algorithm are described in Section II. The performance of the proposed method is compared with the existing ones.

Notation: Vectors and matrices are denoted by boldface lowercase and uppercase letters, respectively. The transpose, complex conjugate, and conjugate transpose are represented by $(\cdot)^T$, $(\cdot)^*$, and $(\cdot)^H$, respectively. Tr(·) denotes the trace of a matrix. The notation $\|\cdot\|$ stands for the Euclidean norm for vectors. $\mathbb{E}[x]$ represents the expectation of variable x.

II. System Model

We consider a coded M-MIMO orthogonal frequency division multiplexing (OFDM) system, where the BS is equipped with N_R antennas and serves N_T single-antenna users simultaneously. The transmitted bit streams from N_T users are separately encoded by a channel encoder, and then mapped to a sequence of energy-normalized complex-valued quadrature-amplitude-modulated constellation points. To ease presentation, we only consider one single subcarrier. The input-output relation between the transmitted and received signals on subcarrier l is expressed as

$$\mathbf{y}^{(l)} = \mathbf{H}^{(l)}\mathbf{s} + \mathbf{n}^{(l)}, \tag{1}$$

where $\mathbf{y}^{(l)} \in C^{N_R \times 1}$ represents the received vector on subcarrier l, $\mathbf{H}^{(l)} \in C^{N_R \times N_T}$ denotes the spatially multiplexed complex channel matrix, $\mathbf{s} \in C^{N_T \times 1}$ is the transmitted symbol vector, and $\mathbf{n}^{(l)} \in C^{N_R \times 1}$ is the complex additive white Gaussian noise vector with zero mean and variance σ^2 per complex entry. When there is no ambiguity, we will suppress the superscript l in the remainder of this manuscript. The BUAR is defined as $\beta = \frac{N_T}{N_R}$.

The estimate of the transmitted signal vector $\mathbf{s}$ based on the linear MMSE principle is given by

$$\hat{\mathbf{s}} = \left(\mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{H}^H\mathbf{y} = \mathbf{W}^{-1}\mathbf{y}^{\mathrm{MF}} = \mathbf{W}^{-1}\mathbf{G}\mathbf{s} + \mathbf{W}^{-1}\mathbf{n}, \tag{2}$$

where $\mathbf{y}^{\mathrm{MF}} = \mathbf{H}^H\mathbf{y}$ denotes the matched-filter output of $\mathbf{y}$, $\mathbf{W} = \mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I}$ is the LMMSE filtering matrix and $\mathbf{G} = \mathbf{H}^H\mathbf{H}$ represents the Gram matrix. Let $\mathbf{U} = \mathbf{W}^{-1}\mathbf{G} = \mathbf{I} - \sigma^2\mathbf{W}^{-1}$ denote the effective channel matrix and $\mathbf{V} = \mathbf{W}^{-1}\mathbf{G}\mathbf{W}^{-1}$ be the equivalent noise covariance matrix. We decompose the $\hat{\mathbf{s}}$ in (2) for user i as follows:

$$\hat{s}_i = \mathbf{w}_i^H\mathbf{g}_i s_i + \sum_{j\neq i}\mathbf{w}_j^H\mathbf{g}_j s_j + \mathbf{w}_i^H\mathbf{n} = \mu_i s_i + \tilde{n}_i, \tag{3}$$

where $\mathbf{w}_i^H$ and $\mathbf{g}_i$ are the i-th row of $\mathbf{W}^{-1}$ and i-th column of $\mathbf{G}$, respectively, $\mu_i = \mathbf{w}_i^H\mathbf{g}_i$ denotes the equivalent channel gain, and $\tilde{n}_i$ represents the noise-plus-interference (NPI) of user i with variance $\nu_i^2 = \sum_{j\neq i}^{N_T}|\mathbf{V}_{ij}|^2 + \mathbf{V}_{ii}\sigma^2$. The *a posteriori* LLR of bit k for the i-th user can be obtained by the by the max-log approximation as follows:

$$\mathcal{L}_{i,k} = \rho_i\left(\min_{a\in\mathcal{X}_k^0}\left|\frac{\hat{s}_i}{\mu_i} - a\right|^2 - \min_{a'\in\mathcal{X}_k^1}\left|\frac{\hat{s}_i}{\mu_i} - a'\right|^2\right), \tag{4}$$

where $\rho_i = \mu_i^2/\nu_i^2$ denotes the post-equalization signal-to-interference-and-noise-ratio (SINR), $\mathcal{X}_k^{(0)}$ and $\mathcal{X}_k^{(1)}$ are the sets containing the symbols from the modulation constellation $\mathcal{X}$ with the k-th bit being 0 and 1, respectively.

The LMMSE algorithm requires matrix inversion $\mathbf{W}^{-1}$ to obtain the LMMSE estimate $\hat{\mathbf{s}}$ [cf. (2)], the effective channel gain $\mathbf{U} = \mathbf{W}^{-1}\mathbf{G}$, and the NPI variance to calculate the approximate LLRs for soft-input channel decoding. The computation of $\mathbf{W}^{-1}$ requires $\mathcal{O}(N_T^3)$ number of operations [7], which results in prohibitive complexity for a M-MIMO system with a large number of users.

III. Proposed Joint WNS-GS Soft-Output Signal Detection

A. Signal Detection based on the Joint WNS and GS Method

Since the LMMSE filtering matrix $\mathbf{W}$ is Hermitian positive, we can decompose $\mathbf{W}$ as

$$\mathbf{W} = \mathbf{D} + \mathbf{L} + \mathbf{L}^H, \tag{5}$$

where $\mathbf{D}$, $\mathbf{L}$, and $\mathbf{L}^H$ are the diagonal, the strictly lower triangular, and upper triangular parts of $\mathbf{W}$, respectively.

Based on the GS method [6], the transmitted signal vector $\mathbf{s}$ is estimated as follows:

$$\mathbf{s}^{(t)} = (\mathbf{D} + \mathbf{L})^{-1}(\mathbf{y}^{\mathrm{MF}} - \mathbf{L}^H \mathbf{s}^{(t-1)}), \quad t = 1, 2, \cdots, \tag{6}$$

where t denotes the iteration number, and $\mathbf{s}^{(0)}$ represents the initial solution. In general, if there is no *a priori* information about the final solution, the initial solution $\mathbf{s}^{(0)}$ in (6) is normally set to be a zero-vector.

The initial solution $\mathbf{s}^{(0)}$ would impact the convergence rate greatly and affect both complexity and accuracy of the final solution, in particular for complexity constrained practical implementation. An initial solution, which is close to the LMMSE estimate $\hat{\mathbf{s}}$, can lead to a faster convergence than the iterative method which is started at zero-vector. However, to find a good initial $\mathbf{s}^{(0)}$ is as difficult as determining the LMMSE estimate $\hat{\mathbf{s}}$.

To reduce the computational burden of the matrix inversion, the NS expansion [8] is proposed to implement $\mathbf{W}^{-1}$ as follows

$$\mathbf{W}^{-1} = \sum_{n=0}^{\infty} (\mathbf{I}_{N_T} - \Theta^{-1}\mathbf{W})^n \Theta^{-1}, \tag{7}$$

where Θ is an arbitrary matrix satisfying the condition $\lim_{n\to\infty}(\mathrm{I}_{N_T} - \Theta^{-1}\mathbf{W})^n \to \mathbf{0}_{N_T}$. We can decompose $\mathbf{W}$ in (5) into its main diagonal $\mathbf{D}$ and off-diagonal $\mathbf{B}$ such that $\mathbf{W} = \mathbf{D} + \mathbf{B}$. Since $\mathbf{W}$ is a diagonally dominant matrix, we can choose $\Theta = \mathbf{D}$ to save computational complexity. By keeping only the first L terms of the Neumann series, we obtain the L-term approximation given by:

$$\mathbf{W}_L^{-1} = \sum_{n=0}^{L-1} (-\mathbf{D}^{-1}\mathbf{B})^n \mathbf{D}^{-1}. \tag{8}$$

The number of expansion terms, L, can be used as a tuning parameter to trade-off between complexity and accuracy. For $L = 1$, the NS-based approach degenerates into the matched-filter detector. For $L = 2$, the inverse of $\mathbf{W}$ is approximated by

$$\mathbf{W}_2^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}, \tag{9}$$

which requires $\mathcal{O}(N_T^2)$ number of operations and is significantly lower than $\mathcal{O}(N_T^3)$ operations

required by an exact inversion approach.

For $L > 3$, the NS-based approach entails computational complexity $O(N_T^3)$ which is comparable to or even higher than that of the exact matrix inversion [4]. For a system with a small BUAR, a relatively large number expansion term L (e.g., 4) is usually required to achieve a satisfactory performance [6]. Therefore, the NS-based approach with $L > 3$ may not be a cost-effective option for LMMSE detection [6]. Nonetheless, it is reasonable to assume that an NS-based approach can provide a promising searching direction *even* for a limited number of expansion terms. Based on this observation, we propose to utilize the two-term NS-based approach of (9) to obtain the initial solution given by:

$$\mathbf{s}^{(0)} = \mathbf{W}_2^{-1}\mathbf{y}^{\mathrm{MF}}. \tag{10}$$

Since the NS with a limited number of expansion terms will inevitably introduce approximation errors, we can introduce a weighted Neumann series (WNS) to mitigate it as follows:

$$\mathbf{W}^{-1} \approx \widehat{\mathbf{W}}_L^{-1} = \sum\nolimits_{n=0}^{L-1} \alpha_{n-1}(\mathbf{I}_{N_T} - \mathbf{D}^{-1}\mathbf{B})^{n-1}\mathbf{D}^{-1}, \tag{11}$$

where $\alpha \triangleq [\alpha_0, \cdots, \alpha_{L-1}]^{\mathrm{T}}$ are the weights to provide better approximation to the exact matrix inversion. The objective is to minimize the approximation error (AE) between the exact matrix inversion $\mathbf{W}^{-1}$ and the approximate matrix inversion based on the WNS of (11) given by

$$\arg\min_{\alpha\in\mathbb{C}^L} AE = \arg\min_{\alpha\in\mathbb{C}^L} \mathbb{E}\left[\left\|(\mathbf{W}\widehat{\mathbf{W}}_L^{-1} - \mathbf{I}_{N_T})/N_T\right\|_{\mathrm{F}}^2\right], \tag{12}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm.

Since $\mathbf{W}$ and $\widehat{\mathbf{W}}_L^{-1}$ are Hermitian, we can constrain α to be a real-valued number without loss of optimality. For a 2-term WNS, we can utilize a numerical method to obtain optimized real-valued α_0 and α_1. We will elaborate the determination of the optimal α_0 and α_1 later in Section IV-A.

B. LLR Computation based on the Proposed WNS

Since we utilize the 2-term WNS to approximate the exact inverse $\mathbf{W}^{-1}$, the approximate equivalent channel matrix is given by

$$\hat{\mathbf{U}} = \widehat{\mathbf{W}}_2^{-1}\mathbf{G} \approx \mathbf{I}_{N_T} - \sigma^2\widehat{\mathbf{W}}_2^{-1}. \tag{13}$$

Therefore, the approximate effective channel gain of user i is expressed as

$$\hat{\mu}_i = \hat{\mathbf{U}}_{ii} = 1 - \sigma^2(\widehat{\mathbf{W}}_2^{-1})_{ii}, \tag{14}$$

where $(\widehat{\mathbf{W}}_2^{-1})_{ii}$ denotes the i-th diagonal entry of $\widehat{\mathbf{W}}_2^{-1}$. The effective noise variance percieved from user i is given by $\hat{\upsilon}_i = \hat{\mu}_i(1 - \hat{\mu}_i)$. Based on (4) and (14), we can efficiently obtain the approximate LLR.

C. Computational Complexity Analysis

Since $\mathbf{W}$ and the matched-filter output $\mathbf{y}^{\mathrm{MF}}$ are all required by the conventional LMMSE algorithm and the proposed joint WNS-GS method, we focus on the complexity of LLR computation. We assume that one complex-valued multiplication can be realized by four real-valued multiplications and two real-valued additions. Since the complexity of the additions is much simpler than the multiplications, the evaluation is based on the required number of real-valued multiplications.

1) Computation of $\mathbf{W}_2^{-1}$: Since the multiplication of a $N_T \times N_T$ real-valued diagonal matrix and a $N_T \times N_T$ complex-valued matrix requires $2N_T^2$ real-valued multiplications, the computation of $\mathbf{W}_2^{-1}$ requires $2(2N_T^2 - 2N_T)$ real-valued multiplications based on (9). Since the $\mathbf{W}_2^{-1}$ is Hermitian, the total number of real-valued multiplications is $2N_T^2 - 2N_T$.
2) Computation of $\mathbf{s}^{(0)}$: The matrix-vector multiplication of (10) entails $4N_T^2$ real-valued multiplications.
3) Computation of $\mathbf{s}^{(t)}$: Based on (6), the solution $\mathbf{s}^{(t)}$ can be expressed as

$$s_i^{(t)} = \frac{1}{\mathbf{W}_{ii}} \times \left(y_i^{\mathrm{MF}} - \sum_{j=1}^{i-1} \mathbf{W}_{ij} s_j^{(t)} - \sum_{j=i+1}^{N_T} \mathbf{W}_{ij} s_j^{(t-1)} \right), \tag{15}$$

   where $s_i^{(t)}$, $s_i^{(t-1)}$, y_i^{MF} represent the i-th element of $\mathbf{s}^{(t)}$, $\mathbf{s}^{(t-1)}$, and $\mathbf{y}^{\mathrm{MF}}$, respectively. Based on (15), the computation of $s_i^{(t)}$ requires $4N_T$ real-valued multiplications. Therefore, the $\mathbf{s}^{(t)}$ with dimension $N_T \times 1$ requires $t4N_T^2$ real-valued multiplications.
4) Computation of $\hat{\mu}_i$ and $\hat{\upsilon}_i$: Since the diagonal entries of the $\mathbf{W}_2^{-1}$ are real-valued, it takes N_T real-valued multiplications to obtain μ_i (for $i = 1, \cdots, N_T$) based on (14). The computation of υ_i (for $i = 1, \cdots, N_T$) entails N_T real-valued multiplications.

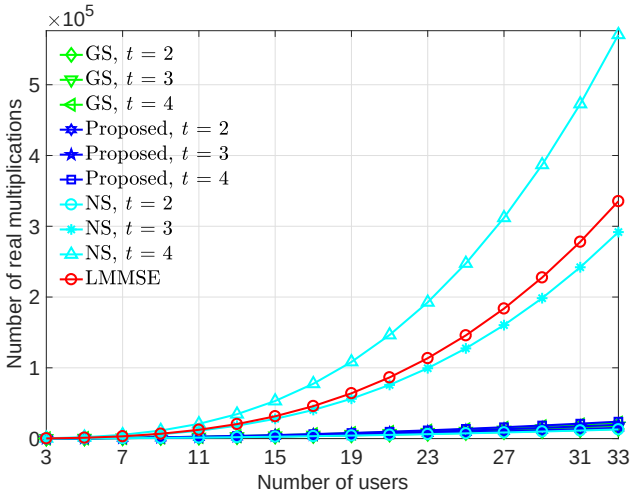

Fig. 1. Complexity comparison of the GS, NSE, LMMSE and the proposed joint WNS-GS method.

To sum up, the overall required number of real-valued multiplications of the proposed joint WNS-GS algorithm with iteration t is $4(t + 3/2)N_T^2$.

Fig. 1 compares the complexity of the NS-based method algorithm [4] and the proposed joint GS-NSE, whereby the LMMSE algorithm based on the Cholesky decomposition is also included as a baseline for comparison. Fig. 1 illustrates that the NS-based method exhibits lower computational complexity than the LMMSE algorithm based on the Cholesky decomposition for $t \leq 3$. The NS-based algorithm loses the complexity advantage over the exact matrix inversion for $t > 4$ as illustrated in Fig. 1.

Fig. 1 shows that the computational complexity of the proposed joint WNS-GS is similar to that of the conventional GS method proposed in [6]. This is explained as follows: The computational overhead of the LLR computation of the GS method [6] is *on par with* that of determination of the initial solution $\mathbf{s}^{(0)}$ [c.f. (10)]. The proposed method requires more effort to obtain the initial solution [c.f. (10)] than that of the diagonal-approximation-based approach [6], and the LLR can be obtained as a *by-product* [c.f. (10) and (14)], thus, entailing comparable computational complexity.

As explained in Section III-A, the proposed method exploits the NS's *exploration* capability to generate a closer-to-LMMSE estimation as the initial solution, therefore, significantly enhancing the subsequent GS's convergence rate. This will be verified later in Section IV-B.

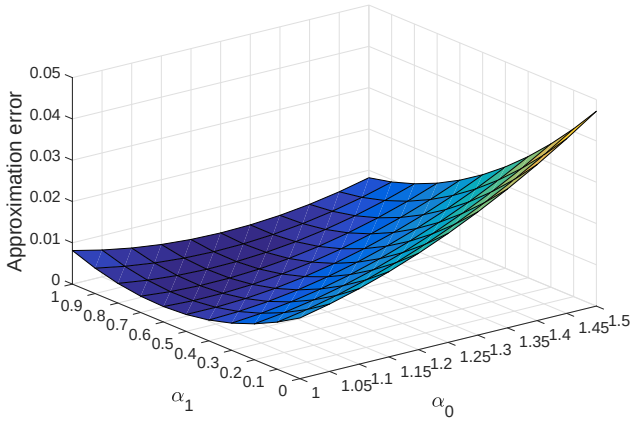

Fig. 2. The AE as a function of the choice of α_0 and α_1 for a 64×24 MIMO system at SNR = 5 dB.

IV. Numerical Results

In this section, we evaluate the block error rate (BLER) performance of the proposed joint WNS-GS method and compare with the conventional NS-based approach [6] and the GS method [4]. A turbo-coded OFDM multiplexing system with 2048-point fast Fourier transform (FFT) and 15-kHz subcarrier spacing is considered in this work. The channel model is the extended vehicular (EVA) model with a maximum Doppler frequency of 10 Hz, which corresponds approximately to a mobile velocity of 5.4 km/h for operation in the 2.0 GHz band. Binary information data bits with a length of 4264 are turbo encoded by a mother code of rate $R_c = 1/3$ (with generator polynomials $[13, 15]_8$) and then punctured to code rate $R_c = 5/6$ as specified in the 3GPP standard [9]. Max-Log-MAP decoding with four decoding iterations was applied in the channel decoder. The modulation scheme of 64 QAM is applied in all simulations.

A. Determination of α_0 and α_1

It is apparent that the determination of optimal α_0 and α_1 is related to the BUAR, SNR, and specific channel statistics. Since the closed-form solution of the α_0 and α_1 is difficult to treat analytically, we use a numerical method to obtain the empirically optimized α_0 and α_1. We first study the relationship between the AE with respect to α_0 and α_1. For 64×24 M-MIMO, Fig. 2 illustrates that the AE reaches the minimum when $\alpha_0 = 1.05$ and $\alpha_1 = 0.55$. The BLER results (not shown due to space limits) illustrate that the proposed method with $\alpha_0 = 1.05$ and $\alpha_1 = 0.55$ achieves the best performance, which validates the observation of the AE. For simplicity, we

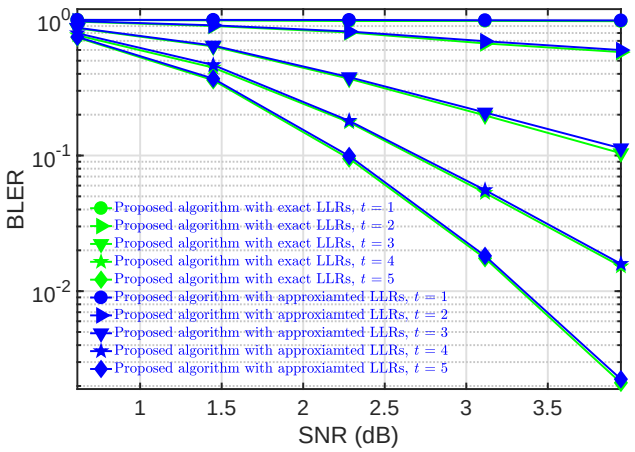

Fig. 3. BLER performance comparison between the exact LLR computation method and the proposed approximation method.

may choose $\alpha_0 = 1.0$ and $\alpha_1 = 0.5$ for low overload (e.g., $\beta < 0.25$), $\alpha_0 = 1.05$ and $\alpha_1 = 0.55$ for intermediate overload (e.g., $0.25 <= \beta < 0.375$), and $\alpha_0 = 1.1$ and $\alpha_1 = 0.6$ for high overload (e.g., $0.375 <= \beta < 0.5$). Overall, the performance degradation due to the mismatch of the optimized weighting factor values for various user loads is marginal based on extensive numerical results.

Remark 1: The results provided in this work are not intended to be exhaustive but rather to serve as a guiding principle. In all cases tested, the proposed joint WNS-GS method exhibits great resilience to the mismatch of the SNR, and/or the BUAR, and/or the channel model, which renders the proposed joint WNS-GS method attractive for implementation in practical systems.

B. BLER Comparison

We first consider the performance of the proposed approximate LLR computation method. Fig. 3 illustrates that the gap between the proposed approximate LLR computation and the exact method is only about 0.05 dB at BLER = 0.2 for $t = 3$. The gap becomes indistinguishable diminish for $t > 3$. Therefore, the proposed approximated LLR computation method can be a low-complexity alternative to the complex exact LLR computation method.

We then compare the proposed method with conventional GS and NS methods [6] in Fig. 4. For iteration $t = 4$, Fig. 4 illustrates that the proposed joint WNS-GS method achieves a performance gain of approximately 1.2 dB at BLER = 10^{-1} compared to the GS method [6] for the 64×24 MIMO (the conventional zero initial solution based results were not depicted

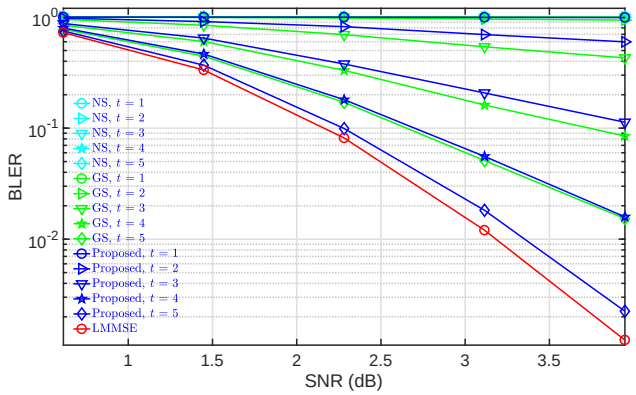

Fig. 4. BLER performance comparison between the NS, GS and the proposed joint WNS-GS method.

since its performance is worse than the diagonal-based one [6]). The NS-based approach does not even converge, i.e., BLER = 1, as depicted in Fig. 4. The gap between the conventional GS method with $t = 4$ and the proposed one with $t = 3$ is negligible. This translates to reduced latency, which is *appealing* for delay-stringent applications. Even more prominent performance gain is observed for a higher loading 64×32 MIMO. Fig. 4 illustrates that the proposed method achieves close-to-LMMSE with a gap of less than 0.2 dB at BLER = 10^{-2} with $t = 4$.

V. Conclusion

In this paper, we proposed a joint WNS-GS approach to implement the LMMSE detection for uplink M-MIMO systems. The proposed method exploits the NS's *exploration* capability to generate a promising initial solution for a quick start of the subsequent GS's iterative detection. A weighted approach is incorporated into the NS to further refine the initial solution, thus significantly enhancing the convergence rate of the GS method. We also proposed to compute approximate LLR based on the WNS approach with negligible performance degradation. Numerical results illustrate that the proposed joint WNS-GS approach outperforms the conventional GS method with comparable computational complexity. The proposed joint WNS-GS method achieves near-LMMSE performance with significantly reduced computational complexity. The extension of the proposed joint WNS-GS paradigm to other signal processing problems involving high-order matrix inversion, such as downlink precoding for M-MIMO systems is straightforward.

VI. Declarations

Not applicable.

References

[1] 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 Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

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

[3] X. Gao, L. Dai, Y. Ma, and Z. Wang, "Low-complexity near-optimal signal detection for uplink large-scale MIMO systems, " *Electro. Lett.*, vol. 50, no. 18, pp. 1326–1328, 2014.

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

[5] M. Wu, B. Yin, G. Wang, C. Dick, J. R. Cavallaro, and C. Studer, "Large-scale MIMO detection for 3GPP LTE: Algorithm and FPGA implementation, " *IEEE J. Sel. Topic of Sig. Proc.*, vol. 8, no. 5, pp. 916–929, Oct. 2014.

[6] L. Dai, X. Gao, X. Su, S. Han, C. L. I, and Z. Wang, "Low-complexity soft-output signal detection based on Gauss-Seidel method for uplink multiuser large-scale MIMO systems, " *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4839–4845, Oct. 2015.

[7] A. Burg, S. Haene, D. Perels, P. Luethi, N. Felber, and W. Fichtner, "Algorithm and VLSI architecture for linear MMSE detection in MIMO OFDM systems, " in *Proc. IEEE ISCAS*, Island of Kos, Greece, May 2006, pp. 4102–4105.

[8] G. Stewart, *Matrix Algorithms*: Basic Decompositions 1998.

[9] "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding, " 3rd Generation Partnership Project (3GPP), 3GPP TS 36.212 V9.1.0, Mar. 2010.