Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5336187, 13 pages
https://doi.org/10.1155/2022/5336187

Research Article

Improved Approximate Expectation Propagation Massive MIMO Detector with Second-Order Richardson Iteration

Qian Deng,[1] **Xuehui Chen**,[1] **Wanting Fu**,[1] **Xiaopeng Liang**,[1,2] **Shilong Xie**,[1] **Feng Shu**,[1] **and Yuan Yuan Wu**[1]

[1]*College of Information and Communication Engineering, Hainan University, Haikou 570228, China*
[2]*Gannan University of Science and Technology, Ganzhou 341000, China*

Correspondence should be addressed to Xiaopeng Liang; liangxiaopeng315@163.com

Received 16 September 2021; Accepted 6 January 2022; Published 16 February 2022

Academic Editor: Andrej Hrovat

The expectation propagation (EP) detector achieves significantly better performance than the linear detectors (such as minimum mean squared error detector) in massive MIMO systems, which has drawn great attention recently. EP's approximation (EPA) algorithm simplifies the update formula of the EP algorithm by reexpressing the moment matching condition so that the number of matrix inversions in the EP algorithm is reduced to one. However, the expense is that the EPA algorithm requires higher accuracy for this inversion; otherwise, the bit-error-rate (BER) performance will suffer serious losses. To tackle this issue, the SORI iterative algorithm is introduced to obtain the high-precision result of this inversion to ensure the good BER performance of the EPA algorithm. First, the new expression of the SORI iterative algorithm is derived under the equivalent real-valued system. Second, the improved EPA-SORI algorithm is then introduced by the SORI algorithm, which is used to approximate the initial value of the EPA algorithm under the real-valued system. Finally, by designing the initial solution and the relaxation factor of the EPA-SORI iterative algorithm, the convergence rate can be quickly increased without increasing the complexity. Simulation and complexity results exhibit that in various massive MIMO system configurations, the proposed EPA-SORI algorithm can achieve the same BER performance as the Exact EP algorithm with significantly lower complexity. At the same time, compared with MMSE and the existing EPA algorithms, the proposed EPA-SORI algorithm has a better performance-complexity trade-off advantage, which is more obvious in scenarios with high modulation order and a large number of users.

1. Introduction

Future wireless communication requires a high data rate and a tremendous amount of connection for emerging applications such as the Internet of things (IoT) [1, 2]. The access of massive IoT devices to the network will lead to tremendous growth in the data volume of mobile communication services, and the wireless network capacity will face unprecedented challenges [3–5]. Aiming at this challenge, main solutions include the usage of larger bandwidth, higher-order MIMO [6, 7], higher-order modulation, more effective coding, and so on. Among these solutions, through deep utilization of spatial dimensions, massive MIMO technology attains enhanced wireless communication capacity and spectral efficiency and has become a key technology for 5 G/B5G wireless communication [8]. With enormous system dimensions and the use of higher-order modulation, signal detection faces a challenge in terms of computational burden and hardware implementation [9].

The traditional optimal signal detector, maximum likelihood (ML) detector, faces the problem of exponential increase in computational complexity for massive MIMO systems [10]. In contrast, linear detection algorithms (such as MMSE, ZF) have reduced computational complexity. Especially when the loading factor $\zeta \ll 1$ ($\zeta \triangleq N_t/N_r$, where N_t and N_r represent the number of single antenna users and the number of base station antennas, respectively), the minimum mean squared error (MMSE), and Zero Forcing

(ZF) linear detection algorithms can achieve near-optimal system performance. However, as a large number of IoT devices are connected to the cellular network, more users need to be served in a cell (e.g., mobile phones, unmanned aerial vehicles (UAVs) [11, 12], sensors [13], vehicle to vehicle (V2V) [14]). Unfortunately, as the number of users increases, the performance of the linear detection algorithm suffers severe degradation. Compared with ML algorithm, other complex detectors (e.g., belief propagation (BP) [15, 16], approximate message passing (AMP) [17]) can achieve excellent performance. However, the convergence speed of the iterative update in the BP algorithm will decrease, and the update formula during the iteration process becomes more complicated, because massive MIMO contains a large number of ring structures when the number of users increases. To address the above issues, an EP algorithm is proposed [18, 19]. For a random value of ζ and high-order modulation, the EP algorithm not only shows significantly better performance than other algorithms such as BP, AMP and MMSE, but also has great flexibility and strong robustness, which has attracted wide attention. However, during each iteration of EP, it is necessary to perform a full matrix inversion with a complexity as high as $\mathcal{O}(N^3)$. In addition, the huge computational cost makes it difficult to implement on hardware. To address this problem, EP's approximation algorithm (EPA) simplifies the update formula of the EP algorithm by reexpressing the moment matching condition, so that the number of matrix inversions in the EP algorithm is reduced to one [20]. However, the expense is that the EPA algorithm requires higher accuracy for this inversion; otherwise, the bit-error-rate (BER) performance will suffer serious performance losses. Recently, some methods based on matrix polynomial decomposition are proposed to apply to the EPA algorithm (such as EPA-NSA and EPA-wNSA) [9, 20]. When loading factor $\zeta \ll 1$, the EPA-NSA algorithm shows good performance. But with the increase of ζ, the EPA-NSA algorithm shows slow convergence or even nonconvergence, resulting in serious degradation of system performance. Although EPA-wNSA algorithm has an improved performance compared with the EPA-NSA algorithm, the degree of improvement is very limited. In addition, the above-mentioned algorithms like EPA-NSA, EPA-wNSA, etc., in spite of avoiding the direct inversion of the matrix and reducing some complexity, calculation of the Gram matrix with a complexity up to $\mathcal{O}(N_t^2 N_r)$ still needs to be involved to obtain high-precision signal detection. Some famous linear iterative method, such as Gauss-Seidel [21–23], successive over relaxation [24, 25], suffer from calculating the Gram matrix and low parallelism. Therefore, they can not directly be applied to alleviate the computational burden of EP iteration. In order to solve these complex issues, an improved algorithm EPA-SORI is proposed in this paper. The EPA-SORI algorithm introduces the SORI iterative algorithm in EPA to obtain a high-precision result of this one inversion to ensure that the EPA algorithm has good error bit rate performance. The contributions of this paper are as follows:

(1) We first deduce a new expression of the SORI iterative algorithm in the real value system. Then, the SORI algorithm under the real value system is further applied to the EPA algorithm. Furthermore, LLR approximation is provided to enhance the accuracy of the EPA-SORI detector.

(2) According to the random matrix theory, under the real-valued system, promising initial solution and relaxation factors are used to further enhance the convergence rate and accuracy and then reduce the computational complexity. Furthermore, a theoretical analysis of the convergence speed of the proposed EPA-SORI algorithm is presented. Theoretical analysis proves that the convergence speed of the proposed EPA-SORI algorithm is significantly higher than the recently reported EPA-wNSA algorithm.

(3) In the iterative process, the proposed EPA-SORI algorithm not only requires no matrix inversion operations but also avoids direct calculation of the Gram matrix, effectively reducing the complexity of the entire algorithm. At the same time, the proposed EPA-SORI algorithm is high-parallel and hardware-friendly.

(4) Simulation and complexity results show that the bit error rate performance of the EPA-SORI algorithm is much better than MMSE, and the complexity is much lower than MMSE. Compared with existing EPA algorithms (such as EPA-INSA, EPA-wNSA, etc.), the proposed EPA-SORI can achieve performance close to Exact EP with lower complexity and higher convergence rate. Furthermore, with high modulation order and a large number of users, EPA-SORI will show a more obvious performance-complexity trade-off advantage.

1.1. Notation. Matrices and column vectors are represented by uppercase and lowercase boldface letters, respectively. The element in i-th row and j-th column of matrix $\mathbf{A}$ is denoted by $\mathbf{A}_{(i,j)}$. $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^{-1}$, $\|\cdot\|_2$, and $|\cdot|$ denote transpose, conjugate transpose, inversion, $2-$norm and determinant, respectively. Also, the probability distribution of $\mathbf{s}$ is denoted by $p(\mathbf{s})$. The real part and imaginary part are denoted by $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$, respectively. $\mathcal{N}(\mathbf{y}: \mu, \Sigma)$ represents the Gaussian probability distribution with mean μ and variance Σ.

2. System Model

We consider a massive MU-MIMO system in which N_r antennas are deployed at the base station to serve N_t users. $\tilde{\mathbf{s}} = [\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_{N_t}]^T$ is a $(N_t \times 1)$-dimensional transmission signal vector, where $\tilde{s}_i \in \tilde{\Omega}$ and $\tilde{\Omega}$ are the symbol set of the $M-$ QAM constellation (e.g., $M = 16/64/256$, etc.). We assume that $\tilde{\mathbf{H}} \in \mathbb{C}^{N_r \times N_t}$ is a flat Rayleigh fading channel.

The received signal $\tilde{\mathbf{y}} \in \mathbb{C}^{N_r \times 1}$ at the receiver can be modeled by the following:

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

where $\tilde{\mathbf{n}}$ is the additive white Gaussian noise (AWGN), and its elements satisfy a complex Gaussian distribution with mean 0 and variance $\tilde{\sigma}_n^2$. Here, the received signal can be reexpressed by the equivalent real-valued augmented vectors as follows:

$$\begin{bmatrix} Re(\tilde{\mathbf{y}}) \\ Im(\tilde{\mathbf{y}}) \end{bmatrix} = \begin{bmatrix} Re(\tilde{\mathbf{H}}) & -\mathrm{Im}(\tilde{\mathbf{H}}) \\ Im(\tilde{\mathbf{H}}) & Re(\tilde{\mathbf{H}}) \end{bmatrix} \begin{bmatrix} Re(\tilde{\mathbf{s}}) \\ Im(\tilde{\mathbf{s}}) \end{bmatrix} + \begin{bmatrix} Re(\tilde{\mathbf{n}}) \\ Im(\tilde{\mathbf{n}}) \end{bmatrix}, \tag{2}$$

where $\mathbf{y} = \begin{bmatrix} Re(\tilde{\mathbf{y}}) & Im(\tilde{\mathbf{y}}) \end{bmatrix}^T \in \mathbb{R}^{2N_r \times 1}$, $\mathbf{s} = \begin{bmatrix} Re(\tilde{\mathbf{s}}) & Im(\tilde{\mathbf{s}}) \end{bmatrix}^T \in \Omega^{2N_t \times 1}$ and Ω represents the set of real and imaginary parts of the point set on the constellation in $M-$ QAM modulation. $\mathbf{n} = \begin{bmatrix} Re(\tilde{\mathbf{n}}) & Im(\tilde{\mathbf{n}}) \end{bmatrix}^T$, where $\mathbf{n} \in \mathcal{N}(0, \sigma_n^2 \mathbf{I}_{2N_r})$ satisfy a real Gaussian distribution with mean 0 and variance $\sigma_n^2 = (\tilde{\sigma}_n^2/2)$. Thus, (1) can be modeled by $\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}$, where $\mathbf{H} = \begin{bmatrix} Re(\tilde{\mathbf{H}}) & -\mathrm{Im}(\tilde{\mathbf{H}}) \\ \mathrm{Im}(\tilde{\mathbf{H}}) & Re(\tilde{\mathbf{H}}) \end{bmatrix}$.

3. EP and EPA Algorithms

EP algorithm is a reasoning method based on Bayesian inference, which is used to estimate the value of $\mathbf{x}$ under the condition that $\mathbf{y}$ is known when a joint distribution $p_{\mathbf{xy}}$ is given [26–28]. If we use $p_{\mathbf{xy}}$ to estimate $\mathbf{x}$ directly, the complexity will increase exponentially with the dimensions of $\mathbf{x}$ and $\mathbf{y}$, which consume huge hardware resources in massive MIMO scenario. Therefore, an approximate distribution is used to estimate $p_{\mathbf{xy}}$. In model $\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}$, the joint posterior distribution of the $\mathbf{y}$ and $\mathbf{s}$ is $p(\mathbf{s}|\mathbf{y}) \propto \mathcal{N}(\mathbf{y}\colon \mathbf{H}\mathbf{s}, \sigma_n^2 \mathbf{I}_{2N_r}) \cdot p(\mathbf{s})$. To facilitate the analysis, according to [7], when the transmitted symbols are independent of each other, the EP detector uses a nonstandard Gaussian distribution to replace the prior distribution $p(\mathbf{s})$ of each transmitting antenna, that is, $\widehat{p}(s_i) = \exp(-(1/2)\Lambda_i s_i^2 + \gamma_i s_i)$. And then, a posterior distribution $\widehat{p}(\mathbf{s}|\mathbf{y})$ whose distribution satisfies the exponential family approximate. $\widehat{p}(\mathbf{s}|\mathbf{y})$ can be expressed as follows:

$$\begin{aligned} \widehat{p}(\mathbf{s}|\mathbf{y}) &= \mathcal{N}\left(\mathbf{y}\colon \mathbf{H}\mathbf{s}, \sigma_n^2 \mathbf{I}_{2N_r}\right) \prod_{i=1}^{2N_t} \widehat{p}(s_i) \\ &= \mathcal{N}\left(\mathbf{y}\colon \mathbf{H}\mathbf{s}, \sigma_n^2 \mathbf{I}_{2N_r}\right) \prod_{i=1}^{2N_t} \exp\left(-\frac{1}{2}\Lambda_i s_i^2 + \gamma_i s_i\right). \end{aligned} \tag{3}$$

The cumulative multiplication in (3) can be transformed into the following form:

$$\widehat{p}(\mathbf{s}|\mathbf{y}) = \mathcal{N}\left(\mathbf{y}\colon \mathbf{H}\mathbf{s}, \sigma_n^2 \mathbf{I}_{2N_r}\right) \exp\left(-\frac{1}{2}\mathbf{s}^T \mathbf{\Lambda}\mathbf{s} + \boldsymbol{\gamma}^T \mathbf{s}\right), \tag{4}$$

where $\gamma = [\gamma_1, \gamma_2 \ldots, \gamma_{2N_t}]^T \in \mathbb{R}^{2N_t \times 1}$ is a real-valued column vector, and $\Lambda = \mathrm{diag}([\Lambda_1, \Lambda_2 \ldots, \Lambda_{2N_t}])$ is a diagonal matrix. According to (4), the expression of mean μ and variance Σ of $\widehat{p}(\mathbf{s}|\mathbf{y})$ is as follows:

$$\mathbf{\Sigma} = \left(\frac{1}{\sigma_n^2}\mathbf{H}^{\mathbf{T}}\mathbf{H} + \mathbf{\Lambda}\right)^{-1}, \tag{5}$$

$$\boldsymbol{\mu} = \mathbf{\Sigma}\left(\frac{1}{\sigma_n^2}\mathbf{H}^T\mathbf{y} + \boldsymbol{\gamma}\right). \tag{6}$$

Having constructed $\widehat{p}(\mathbf{s}|\mathbf{y})$, it is necessary to use the moment matching technique so that $\widehat{p}(\mathbf{s}|\mathbf{y})$ and $p(\mathbf{s}|\mathbf{y})$ are as similar as possible. However, the complexity of $\widehat{p}(\mathbf{s}|\mathbf{y})$ is very high, and it is difficult to directly obtain Λ and γ. Aiming at this problem, a sequential EP algorithm is proposed in [9, 20]. It is assumed that $\widehat{p}(s_i|\mathbf{y})$ is the $i-$ th edge of $\widehat{p}(\mathbf{s}|\mathbf{y})$:

$$\widehat{p}^{\backslash i}(s_i|\mathbf{y}) = \frac{\widehat{p}(s_i|\mathbf{y})}{\exp\left(-0.5\Lambda_i s_i^2 + \gamma_i s_i\right)}. \tag{7}$$

$\gamma_i^0 = 0$ and $\Lambda_i^0 = E_s$ are adopted as the initial solutions of (7), where E_s denotes the mean symbol energy. Then, the alternative distribution expression is expressed as follows:

$$p_r^{(l)}(s_i|\mathbf{y}) = \widehat{p}^{(l)/i}(s_i|\mathbf{y}) \cdot p(s_i), \tag{8}$$

where $\widehat{p}^{(l)/i}(s_i)$ is expressed as follows:

$$\widehat{p}^{(l)/i}(s_i) \propto \exp\left(-0.5V_i s_i^2 + \rho_i s_i\right). \tag{9}$$

Then, the first-order moments η_i^l and the second-order moments χ_i^l of $p_r^{(l)}(s_i|\mathbf{y})$ can be obtained by the following:

$$\eta_i^l = \mathbb{E}_{p_r^{(l)}(s_i|\mathbf{y})}[s_i] = \frac{\sum_{s_i \in \Omega} s_i \exp\left(-0.5V_i^{(l-1)} s_i^2 + \rho_i^{(l-1)} s_i\right)}{\sum_{u \in \Omega} \exp\left(-0.5V_i^{(l-1)} u^2 + \rho_i^{(l-1)} u\right)}, \tag{10}$$

$$\begin{aligned} \chi_i^l &= \mathbb{E}_{p_r^{(l)}(s_i|\mathbf{y})}\left[s_i^2\right] - \mathbb{E}^2_{p_r^{(l)}(s_i|\mathbf{y})}[s_i] \\ &= \frac{\sum_{s_i \in \Omega} s_i \exp\left(-0.5V_i^{(l-1)} s_i^2 + \rho_i^{(l-1)} s_i\right)}{\sum_{u \in \Omega} \exp\left(-0.5V_i^{(l-1)} u^2 + \rho_i^{(l-1)} u\right)} \\ &\quad - \left[\frac{\sum_{s_i \in \Omega} s_i \exp\left(-0.5V_i^{(l-1)} s_i^2 + \rho_i^{(l-1)} s_i\right)}{\sum_{u \in \Omega} \exp\left(-0.5V_i^{(l-1)} u^2 + \rho_i^{(l-1)} u\right)}\right]^2. \end{aligned} \tag{11}$$

Assuming that t_i^l, $h_i^{2(l)}$ are the mean and variance of $q^{(l)/i}(s_i)$, the values of V_i^l and ρ_i^l for l iterations can be obtained as follows:

$$V_i^l = \frac{1}{h_i^{2(l)}} = \frac{1 - \sum_{ii}^{(l)} E_s}{\sum_{ii}^{(l)}}, \tag{12}$$

$$\rho_i^l = \frac{t_i^l}{h_i^{2(l)}} = \frac{\mu_i^{(l)}}{\sum_{ii}^{(l)}}. \tag{13}$$

Next, the updated Λ and γ can be obtained by matching the first-order and second-order moments of $p_r^{(l)}(s_i|\mathbf{y})$ and the cavity marginal distribution $\widehat{p}^{\backslash i}(s_i|\mathbf{y})$. Since the calculation process of the EP algorithm is too cumbersome, and each iteration involves the matrix inversion calculation in equation (5), which brings difficulties to the signal detection in massive MIMO scenarios. To simplify the calculation

process and reduce computational complexity, EP's approximate algorithm EPA is proposed in [20]. The EPA algorithm reexpresses the moment matching conditions and no longer uses explicit matrix matching to calculate the results of each iteration. It simplifies the update formula in the iteration process of the EP algorithm and uses a fixed matrix to estimate the matrix $\mathbf{V}$ before the iteration. Then, the approximate value is used to estimate the value of Σ, and finally, we will obtain an approximate algorithm EPA which is different from the Exact EP iteration process. The EPA algorithm reduces the EP algorithm that requires multiple matrix inversion calculations to one time, which can effectively avoid the impact of multiple inaccurate approximate matrix inversion results. However, the expense is that the EPA algorithm requires a complete and high-precision matrix inversion to obtain an initial value of iteration. The implementation of EPA is detailed in Algorithm 1.

4. The Proposed EPA-SORI Detector

MMSE solution is usually used as the iterative initial solution of EP detection [20], then we also use MMSE solution as the iterative initial solution $\mathbf{t}_0$ of the proposed EPA-SORI algorithm. Define $\mathbf{W} \triangleq \mathbf{H}^T\mathbf{H} + (\sigma_n^2/E_s)\mathbf{I}_{2N_t}$, and the initial iteration $\mathbf{t}_0$ can be obtained by the following:

$$\mathbf{t}_0 = \boldsymbol{\Sigma}^{(0)}\boldsymbol{\mu}^{(0)} = \mathbf{W}^{-1}\mathbf{y}^{MF}. \tag{14}$$

From (14), the initial iteration solution problem can be treated as a solution of the equation: $\mathbf{W}\mathbf{t}_0 = \mathbf{y}^{MF}$. Note that the result of this equation is required to be highly accurate; otherwise, the bit error rate performance will suffer serious degradation, so we propose to apply SORI iteration to obtain a high-precision iterative initial solution. In the complex value system, the corresponding SORI algorithm can be constructed as follows:

$$\overline{\mathbf{s}}^{l+1} = \overline{\mathbf{s}}^{l-1} + \alpha_1\beta_1\left(\widetilde{\mathbf{y}}^{MF} - \widetilde{\mathbf{W}}\right)\overline{\mathbf{s}}^{l} + \alpha_1\left(\overline{\mathbf{s}}^{l} - \overline{\mathbf{s}}^{l-1}\right), \tag{15}$$

where $l \geq 1$, $\overline{\mathbf{s}}$ denotes the estimated value of $\widetilde{\mathbf{W}}^{-1}\widetilde{\mathbf{y}}^{MF}$, and α_1 is the relaxation factor, $\beta_1 = 2/(\widetilde{\lambda}_{\max} + \widetilde{\lambda}_{\min})$, $\widetilde{\lambda}_{\max}$ and $\widetilde{\lambda}_{\min}$ are the maximum and minimum eigenvalues of $\widetilde{\mathbf{W}}$, respectively, where $\widetilde{\mathbf{W}} = \widetilde{\mathbf{H}}^H\widetilde{\mathbf{H}} + (\overline{\sigma}_n^2/E_s)\mathbf{I}_{N_t}$. Note that the EPA algorithm is implemented in a real-valued system; thus, we need to apply the SORI algorithm to a real-valued system. Here, the SORI method can be reexpressed as follows:

$$\widehat{\mathbf{s}}^{l+1} = \widehat{\mathbf{s}}^{l-1} + \alpha_2\beta_2\left(\mathbf{y}^{MF} - \mathbf{W}\right)\widehat{\mathbf{s}}^{l} + \alpha_2\left(\widehat{\mathbf{s}}^{l} - \widehat{\mathbf{s}}^{l-1}\right), \tag{16}$$

where $l \geq 1$, $\widehat{\mathbf{s}} \in \mathbb{R}^{2N_r \times 1}$ denotes the estimated value of $\mathbf{t}_0$, and α_2 denotes the relaxation factor, $\beta_2 = 2/(\widetilde{\lambda}_{\max} + \widetilde{\lambda}_{\min})$, and $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum eigenvalue and minimum eigenvalue of $\mathbf{W}$, respectively. The relationship between $\mathbf{y}^{MF}$ and $\widetilde{\mathbf{y}}^{MF}$ satisfies: $\mathbf{y}^{MF} = \left[\, Re(\widetilde{\mathbf{y}}^{MF}) \;\; Im(\widetilde{\mathbf{y}}^{MF}) \,\right]^T$. In recently reported literature [29], the SORI algorithm is mainly applied to signal detection under complex-valued systems and cannot be directly applied to real-valued systems. Thus, it is necessary to further deduce the maximum and minimum eigenvalues of $\mathbf{W}$. As long as the eigenvalues of $\mathbf{H}^T\mathbf{H}$ can be obtained, the eigenvalues of $\mathbf{W}$ can be obtained.

Lemma 1. *In the real-valued system, the maximum and minimum eigenvalues of* $\mathbf{W}$ *are as follows:*

$$\begin{cases} \lambda_{\max} \longrightarrow N_{\mathrm{r}}\left(1 + \sqrt{\dfrac{Size_C(\mathbf{H})}{Size_R(\mathbf{H})}}\right)^2 \\ \\ \lambda_{\min} \longrightarrow N_r\left(1 - \sqrt{\dfrac{Size_C(\mathbf{H})}{Size_R(\mathbf{H})}}\right)^2 \end{cases}, \tag{17}$$

where $Size_C(\cdot)$ *and* $Size_R(\cdot)$ *denote the number of columns and rows.*

Proof. Assuming that one of the eigenvalues of $\mathbf{H}^T\mathbf{H}$ is λ_1, and one of the eigenvalues of $\widetilde{\mathbf{H}}^H\widetilde{\mathbf{H}}$ is λ. Based on the properties of eigenvalues, we have the following:

$$\left|\lambda\mathbf{I}_{N_t} - \widetilde{\mathbf{H}}^H\widetilde{\mathbf{H}}\right| = 0, \tag{18}$$

$$\left|\lambda_1\mathbf{I}_{2N_t} - \mathbf{H}^T\mathbf{H}\right| = 0. \tag{19}$$

Here, $\widetilde{\mathbf{H}}^H\widetilde{\mathbf{H}}$ and $\mathbf{H}^T\mathbf{H}$ can be expressed as follows:

$$\widetilde{\mathbf{H}}^H\widetilde{\mathbf{H}} = Re^2(\widetilde{\mathbf{H}}) + Im^2(\widetilde{\mathbf{H}}), \tag{20}$$

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} Re^2(\widetilde{\mathbf{H}}) + Im^2(\widetilde{\mathbf{H}}) & 0 \\ 0 & Re^2(\widetilde{\mathbf{H}}) + Im^2(\widetilde{\mathbf{H}}) \end{bmatrix}. \tag{21}$$

Substitute (20) and (21) into (18) and (19), and we have the following:

$$\left|\lambda\mathbf{I}_{N_t} - \left[Re^2(\widetilde{\mathbf{H}}) + Im^2(\widetilde{\mathbf{H}})\right]\right| = 0, \tag{22}$$

$$\left|\begin{bmatrix} \lambda_1\mathbf{I}_{N_t} - \widetilde{\mathbf{H}}^H\widetilde{\mathbf{H}} & 0 \\ 0 & \lambda_1\mathbf{I}_{N_t} - \widetilde{\mathbf{H}}^H\widetilde{\mathbf{H}} \end{bmatrix}\right| = 0. \tag{23}$$

Simplify (23), and we have the following:

$$\left|\lambda_1\mathbf{I}_{N_t} - \left[Re^2(\widetilde{\mathbf{H}}) + Im^2(\widetilde{\mathbf{H}})\right]\right|^2 = 0. \tag{24}$$

The comparison between (24) and (22) shows that the eigenvalues of the two are the same, i.e., $\lambda_{\max} \longrightarrow N_{\mathrm{r}}(1 + \sqrt{(N_t/N_r)})^2, \lambda_{\min} \longrightarrow N_r(1 - \sqrt{(N_t/N_r)})^2$. In addition, we can get the same conclusion by using random matrix theory [30, 31]. According to random matrix theory [30], in a real-valued system, $\mathbf{H}_{(i,j)} \sim \mathcal{N}(0, (1/2))$. Thus, $(1/\sqrt{N_r})\mathbf{H}_{(i,j)} \sim \mathcal{N}(0, (1/2N_r))$. Hence, we have $\lambda_{\max}(\mathbf{H}^T\mathbf{H}) = N_r(1 + \sqrt{(2N_t/2N_r)})^2$ and $\lambda_{\min}(\mathbf{H}^T\mathbf{H}) = N_r(1 - \sqrt{(2N_t/2N_r)})^2$. Thus, Lemma 1 is proved.

The next step is to derive the relaxation factor α_2 in the real-valued system. According to [29], $\alpha_2 \triangleq (2/1 + \sqrt{1 - \rho^2(\mathbf{G}_{ri})})$, where $\rho(\mathbf{G}_{ri})$ denotes the spectral radius of $\mathbf{G}_{ri}$, and $\mathbf{G}_{ri} = (\mathbf{I}_{2N_t} - \beta_2\mathbf{W})$. Given that b and a are the maximum and minimum eigenvalues of $\mathbf{G}_{ri}$, $\rho(\mathbf{G}_{ri}) = |b|$, thus α_2 can be obtained by the following formula:

(1) Input: $\mathbf{y}$, $\mathbf{H}$, L, σ_n^2, E_s, ϑ.
(2) Output: $\widehat{\mathbf{s}} = \mathbf{t}^L$.
(3) $\mathbf{y}^{MF} = \mathbf{H}^T\mathbf{y}$.
(4) $\mathbf{W} = \mathbf{H}^T\mathbf{H} + (\sigma_n^2/E_s)\mathbf{I}_{2N_t}$.
(5) $\mathbf{s} = \mathbf{W}^{-1}\mathbf{y}$.
(6) $D_i = \mathbf{h}_i^T\mathbf{h}_i, \quad i = 1, 2, \ldots, 2N_t$.
(7) $\mathbf{D} = \mathrm{diag}(D_1, D_2, \ldots, D_{2N_t})$.
(8) for $i = 1, 2, ..., 2N_t$ do
(9) $t_i^0 = s_i/(1 - D_iE_s)$;
(10) end for
(11) repeat
(12) for $i = 1, 2, \ldots, 2N_t$ do
(13) $\eta_i = \mathrm{argmin}_{u_i \in \Omega}|u_i - t_i^{(l-1)}|^2$;
(14) end for
(15) $\mathbf{h}_l = \sigma_n^{-2}\mathbf{H}\eta$.
(16) $\mathbf{m} = \mathbf{b} - \mathbf{H}^{\mathrm{T}}\mathbf{h}_l$.
(17) $\mathbf{t}^l = (\mathbf{D}^{-1})^T\mathbf{m} + \eta^l$.
(18) $\mathbf{t}^l = (1 - \vartheta)\mathbf{t}^{(l-1)} + \vartheta\mathbf{t}^l$, the damping factor $\vartheta \in [0, 1]$.
(19) $l = l + 1$;
(20) until convergence or $l > L$.

Algorithm 1: The EPA algorithm.

$$\alpha_2 = \frac{2}{1 + \sqrt{1 - b^2}}. \tag{25}$$

Since

$$a = -\beta_2\lambda_{\max} + 1 = \frac{\lambda_{\min} - \lambda_{\max}}{\lambda_{\min} + \lambda_{\max}}, \tag{26}$$

$$b = -\beta_2\lambda_{\min} + 1 = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\min} + \lambda_{\max}}. \tag{27}$$

Thus, we have the following:

$$\alpha_2 = \frac{2}{1 + \sqrt{1 - \left(\lambda_{\max} - \lambda_{\min}/\lambda_{\min} + \lambda_{\max}\right)^2}}. \tag{28}$$

It can be seen that α_2 is only related to $\lambda_{\min}$ and $\lambda_{\min}$.

According to the typical properties of the massive MIMO channel, $\widetilde{\mathbf{H}}$ satisfies asymptotic orthogonality, and the Gram matrix $\widetilde{\mathbf{W}}$ is dominant diagonally. From (20) and (21), we have $\mathbf{W} = \begin{bmatrix} \widetilde{\mathbf{W}} & 0 \\ 0 & \widetilde{\mathbf{W}} \end{bmatrix}$, thus, the matrix $\mathbf{W}$ is also dominant diagonally. Therefore, $\mathbf{W}$ could satisfy the approximation as follows:

$$\mathbf{W}_{(i,j)} \approx \begin{cases} \dfrac{(\lambda_{\max} + \lambda_{\min})}{2}, & i = j, \\ 0, & i \neq j. \end{cases} \tag{29}$$

It can be seen from (29), when $\beta_2 = 2/(\lambda_{\max} + \lambda_{\min})$, $\beta_2\mathbf{y}^{MF}$ can be used to approximate $\mathbf{W}^{-1}\mathbf{y}^{MF}$, the SORI algorithm can choose the initial iteration solution of $\widehat{\mathbf{s}}^0 = \mathbf{0}$, $\widehat{\mathbf{s}}^1 = \beta_2\mathbf{y}^{MF}$. The SORI algorithm is summarized in Algorithm 2.

In the proposed EPA-SORI algorithm, $\widehat{\mathbf{s}}^0 = \mathbf{0}$ and $\widehat{\mathbf{s}}^1 = \mathbf{D}^{-1}\mathbf{y}^{MF}$ are used as the initial iteration solutions to further improve the convergence speed. Since $\mathbf{D}$ is the diagonal matrix of the $\mathbf{W}$ matrix, the complexity will not increase. The EPA-SORI algorithm is summarized in Algorithm 3, and an intuitive diagram of the processing of the EPA-SORI algorithm is presented in Figure 1. □

5. Convergence Performance Analysis

This section mainly analyzes the convergence performance of the EPA-SORI algorithm and compares it with that of the EPA-wNSA algorithm. Note that the convergence performance of the EPA-SORI and EPA-wNSA algorithms mainly depends on the convergence performance of the initialization part (i.e., SORI and wNSA).

Lemma 2. *The iterative spectral radius of the proposed EPA-SORI and the EPA-wNSA satisfy the following relationship:*

$$\rho\left(\mathbf{G}_{sori}\right) - \rho\left(\mathbf{G}_w\right) = \left[1 + \delta\left(\sqrt{\zeta} - 1\right)\right]\left(\sqrt{\zeta} - 1\right) < 0. \tag{30}$$

Proof. First, ε^l is denoted as the SORI estimation error after l iterations:

$$\boldsymbol{\varepsilon}^l = \widehat{\mathbf{s}} - \widehat{\mathbf{s}}^l. \tag{31}$$

From (16), the relationship between $\begin{bmatrix} \varepsilon^l \\ \varepsilon^{l+1} \end{bmatrix}$ and $\begin{bmatrix} \varepsilon^{l-1} \\ \varepsilon^l \end{bmatrix}$ can be obtained by the following:

$$\begin{bmatrix} \boldsymbol{\varepsilon}^l \\ \boldsymbol{\varepsilon}^{l+1} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{I}_{2N_t} \\ (1 - \alpha_2)\mathbf{I}_{2N_t} & \alpha_2\mathbf{G}_{ri} \end{bmatrix}\begin{bmatrix} \boldsymbol{\varepsilon}^{l-1} \\ \boldsymbol{\varepsilon}^l \end{bmatrix}. \tag{32}$$

Thus, $\mathbf{G}_{sori}$ denotes the iterative matrix of the SORI iterative algorithm which can be expressed as follows [29]:

(1) Input: $\mathbf{y}$, $\mathbf{H}$, K, σ_n^2,E_s.
(2) Output: $\widehat{\mathbf{s}}^{K+1}$.
(3) $\mathbf{y}^{MF} = \mathbf{H}^T\mathbf{y}$
(4) $\lambda_{\max} \longrightarrow N_{\mathrm{r}}(1+\sqrt{(N_t/N_r)})^2, \lambda_{\min} \longrightarrow N_r(1-\sqrt{(N_t/N_r)})^2$.
(5) $\beta_2 = 2/(\lambda_{\max}+\lambda_{\min})$.
(6) $\alpha_2 = 2/(1+\sqrt{1-[(\lambda_{\max}-\lambda_{\min})/(\lambda_{\min}+\lambda_{\max})]^2})$.
(7) $\widehat{\mathbf{s}}^0 = \mathbf{0}$, $\widehat{\mathbf{s}}^1 = \beta_2\mathbf{y}^{MF}$.
(8) for $l = 1, 2, \ldots, K$ do
(9) $\mathbf{h}_k = \mathbf{H}\widehat{\mathbf{s}}^l$;
(10) $\widehat{\mathbf{s}}^{l+1} = \widehat{\mathbf{s}}^{l-1} + \alpha_2\beta_2(\mathbf{y}^{MF} - \mathbf{H}^T\mathbf{h}_k - \sigma_n^2E_s^{-1}\widehat{\mathbf{s}}^l) + \alpha_2(\widehat{\mathbf{s}}^l - \widehat{\mathbf{s}}^{l-1})$.
(11) end for

Algorithm 2: The SORI algorithm.

(1) Input: $\mathbf{y}$, $\mathbf{H}$, L, K, σ_n^2, E_s, ϑ.
(2) Output: $\mathbf{t}^L$.
(3) $\mathbf{y}^{MF} = \mathbf{H}^T\mathbf{y}$
(4) $\mathbf{b} = \sigma_n^{-2}\mathbf{y}^{MF}$.
(5) $D_i = \mathbf{h}_i^T\mathbf{h}_i, \quad i = 1, 2, \ldots, 2N_t$.
(6) $\mathbf{D} = \mathrm{diag}(D_1, D_2, \ldots, D_{2N_t})$.
(7) $\lambda_{\max} \longrightarrow N_{\mathrm{r}}(1+\sqrt{(N_t/N_r)})^2, \lambda_{\min} \longrightarrow N_r(1-\sqrt{(N_t/N_r)})^2$.
(8) $\beta_2 = 2/(\lambda_{\max}+\lambda_{\min})$
(9) $\alpha_2 = 2/(1+\sqrt{1-[(\lambda_{\max}-\lambda_{\min})/(\lambda_{\min}+\lambda_{\max})]^2})$
(10) $\widehat{\mathbf{s}}^0 = 0$, $\widehat{\mathbf{s}}^1 = \mathbf{D}^{-1}\mathbf{y}^{MF}$.
(11) for $l = 1, 2, ..., K$ do
(12) $\mathbf{h}_k = \mathbf{H}\widehat{\mathbf{s}}^l$;
(13) $\widehat{\mathbf{s}}^{l+1} = \widehat{\mathbf{s}}^{l-1} + \alpha_2\beta_2(\mathbf{y}^{MF} - \mathbf{H}^T\mathbf{h}_k - \sigma_n^2E_s^{-1}\widehat{\mathbf{s}}^l) + \alpha_2(\widehat{\mathbf{s}}^l - \widehat{\mathbf{s}}^{l-1})$.
(14) end for
(15) $\mathbf{s} = \widehat{\mathbf{s}}^{K+1}, \Gamma^2 = \mathrm{diag}(D_1^{-1}, D_2^{-1}, \ldots, D_{2N_t}^{-1})$
(16) for $i = 1, 2, \ldots, 2N_t$ do
(17) $t_i^0 = s_i/(1-\Gamma_i^2E_s)$
(18) end for
(19) repeat
(20) for $i = 1, 2, \ldots, 2N_t$ do
(21) $\eta_i = \mathrm{argmin}_{u_i\in\Omega}|u_i - t_i^{(l-1)}|^2$
(22) end for
(23) $\mathbf{h}_l = \sigma_n^{-2}\mathbf{H}\eta$
(24) $\mathbf{m} = \mathbf{b} - \mathbf{H}^{\mathrm{T}}\mathbf{h}_l$
(25) $\mathbf{t}^l = (\Gamma^2)^T\mathbf{m} + \eta^l$
(26) $\mathbf{t}^l = (1-\vartheta)\mathbf{t}^{(l-1)} + \vartheta\mathbf{t}^l$, the damping factor $\vartheta \in [0, 1]$.
(27) $l = l + 1$.
(28) until convergence or $l > L$.

Algorithm 3: Proposed EPA-SORI algorithm.

$$\mathbf{G}_{sori} = \begin{bmatrix} 0 & \mathbf{I}_{2N_t} \\ (1-\alpha_2)\mathbf{I}_{2N_t} & \alpha_2\mathbf{G}_{ri} \end{bmatrix}. \tag{33}$$

Assuming that λ_{ri} and λ_{sori} are the eigenvalues of $\mathbf{G}_{ri}$ and $\mathbf{G}_{sori}$, respectively. b_1 and a_1 are the maximum and minimum eigenvalues of $\mathbf{G}_{sori}$, respectively. At the same time, b and a are the maximum and minimum eigenvalues of $\mathbf{G}_{ri}$, respectively. Thus, we have $|\lambda_{ri}\mathbf{I}_{2N_t} - \mathbf{G}_{ri}| = 0$ and $|\lambda_{sori}\mathbf{I}_{4N_t} - \mathbf{G}_{sori}| = 0$. Then, after substituting (33) into $|\lambda_{sori}\mathbf{I}_{4N_t} - \mathbf{G}_{sori}| = 0$, we have $|(\lambda_{sori}^2 + \alpha_2 - 1/\lambda_{sori}\alpha_2)$

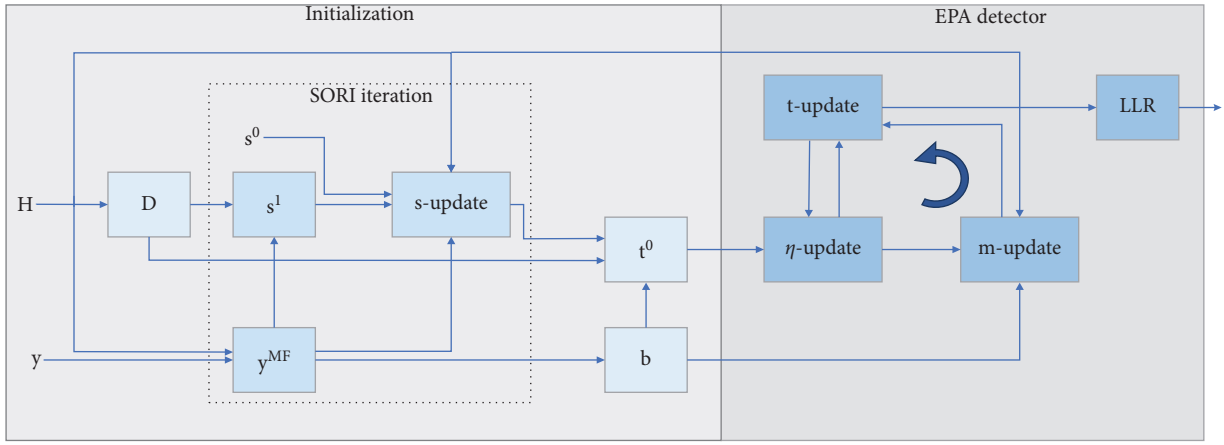

Figure 1: Functional diagram of EPA-SORI processing and detector.

$\mathbf{I}_{2N_t} - \mathbf{G}_{ri}| = 0$. Thus, compared with $|\lambda_{ri}\mathbf{I}_{2N_t} - \mathbf{G}_{ri}| = 0$, it can be derived as follows:

$$\lambda_{sori}^2 - (\lambda_{ri}\lambda_{sori} - 1)\alpha_2 - 1 = 0. \tag{34}$$

Next, by substituting b into (34), the maximum eigenvalue b_1 of $\mathbf{G}_{sori}$ can be obtained by $b_1 = (\alpha_2 b \pm \sqrt{m}/2)$, where $m = (\alpha_2 b)^2 - 4(\alpha_2 - 1)$. And since $\alpha_2 = (2/1 + \sqrt{1 - b^2})$, m can be calculated as follows:

$$\begin{aligned}(\alpha_2 b)^2 - 4(\alpha_2 - 1) &= \frac{4b^2}{\left(1 + \sqrt{1 - b^2}\right)^2} - \frac{4\left(1 - \sqrt{1 - b^2}\right)}{1 + \sqrt{1 - b^2}} \\ &= \frac{4b^2 - 4\left(1 + \sqrt{1 - b^2}\right)\left(1 - \sqrt{1 - b^2}\right)}{\left(1 + \sqrt{1 - b^2}\right)^2} = 0.\end{aligned} \tag{35}$$

Based on the above derivation, we have $b_1 = (\alpha_2 b/2)$. And after simplification, the spectrum radius of $\mathbf{G}_{sori}$ can be calculated by the following:

$$\rho(\mathbf{G}_{sori}) = b_1 = \sqrt{\alpha_2 - 1} \approx \sqrt{\frac{N_t}{N_r}} = \sqrt{\zeta}. \tag{36}$$

Assuming that $\mathbf{G}_w$ denotes the iterative matrix of the wNSA iterative algorithm in a real-valued system, which is given by the following:

$$\rho(\mathbf{G}_w) = 1 - \delta\rho\left(\mathbf{D}^{-1}\mathbf{W}\right), \tag{37}$$

where δ denotes the weight factor of EPA-wNSA, and $\mathbf{D}$ satisfies the following approximation:

$$\mathbf{D}_{(i,j)} \approx \begin{cases} N_r + \dfrac{\sigma_n^2}{E_s}, & i = j, \\ 0, & i \neq j. \end{cases} \tag{38}$$

Thus, we have the following:

$$\rho(\mathbf{G}_w) = 1 - \delta\frac{\rho(\mathbf{W})}{N_r + \left(\sigma_n^2/E_s\right)}. \tag{39}$$

(39) can be further simplified by substituting the minimum eigenvalue $\lambda_{\min}$ of $\mathbf{W}$. Furthermore, (σ_n^2/E_s) can be neglected since it is much smaller than N_r. Thus, we have the following:

$$\rho(\mathbf{G}_w) = 1 - \delta(1 - \sqrt{\zeta})^2. \tag{40}$$

Comparing the spectrum radius $\mathbf{G}_{sori}$ and $\mathbf{G}_w$, we have the following:

$$\rho(\mathbf{G}_{sori}) - \rho(\mathbf{G}_w) = [1 + \delta(\sqrt{\zeta} - 1)](\sqrt{\zeta} - 1), \tag{41}$$

where $0 < \delta < 1, 0 < \zeta < 1$; hence $[1 + \delta(\sqrt{\zeta} - 1)](\sqrt{\zeta} - 1) < 0$. Thus, Lemma 2 is proved. The convergence rate R of the iterative algorithm is closely related to the spectral radius of the iterative matrix $\mathbf{G}$, i.e., $R = -\lg(\rho(\mathbf{G}))$. Hence, the smaller the spectral radius of the iteration matrix, the greater the convergence speed. From (30), $\mathbf{G}_{sori}$ has a smaller spectrum radius than $\mathbf{G}_w$; it means that the proposed EPA-SORI algorithm exhibits favorable convergence performance. □

6. Computational Complexity Analysis

In this section, the computational complexity is given by the number of real-valued multiplications (RMULs). As shown in Figure 1, the entire calculation process is divided into two parts: the initialization process and the iterative processes of the EPA-SORI algorithm.

Firstly, in the initialization process, the number of RMULs required by $\mathbf{y}^{MF}$ and $\mathbf{b}$ are $4N_tN_r$ and $2N_t$, respectively. Since $\mathbf{D}$ is a diagonal matrix, thus the complexity is $4N_tN_r$. Please note here that compared with matrix multiplication, the complexity of $\lambda_{\max}$, $\lambda_{\min}$ and α_2 can be ignored. Thus, the initial solution $\hat{\mathbf{s}}^0$ does not need to be calculated, and the number of RMULs involved in calculating $\hat{\mathbf{s}}^1$ is $4N_t$. The next step is the SORI iteration part of the initialization process. The RMULs of $\mathbf{h}_k$ and $\hat{\mathbf{s}}^{l+1}$ for each

iteration is approximately $4N_tN_r$ and $4N_tN_r + 6N_t$, respectively. Therefore, the RMULs complexity involved in calculating the SORI iteration is $K(8N_tN_r + 6N_t)$, where K denotes the number of iterations of SORI. The final step of this process is the initialization of $\mathbf{t}$, whose computational complexity is $4N_t$.

Secondly, the RMULs complexity involved in calculating the EPA iteration is $L(8N_tN_r + 8N_t)$, where L denotes the number of EPA iterations.

Finally, the overall complexity of the proposed EPA-SORI algorithm is approximately as follows:

$$(K+1)(8N_tN_r + 6N_t) + L(8N_tN_r + 8N_t) + 4N_t. \quad (42)$$

Table 1 shows the computational complexity of EPA-SORI algorithm, MMSE [10], EPA-INSA [9], EPA-wNSA [32], and the Exact EP [18] algorithm. From Table 1, note that Gram matrix calculations with a complexity of up to $4N_t^2N_r$ cannot be avoided in many of the reported methods, such as MMSE, EPA-wNSA, and the Exact EP. Additionally, MMSE and the Exact EP also involve the inversion of a matrix with a complexity of up to $4N_t^3$. In contrast, the proposed EPA-SORI algorithm only involves operations with complexity of about $8N_tN_r(K+1) + 8N_tN_rL$. This is because in the SORI algorithm, direct calculation of the Gram matrix is avoided by splitting calculation. For example, to calculate $\mathbf{H}^T\mathbf{H}\hat{\mathbf{s}}^1$, we first calculate $\mathbf{H}\hat{\mathbf{s}}^1$, and then multiply the result with $\mathbf{H}^T$. Here, we have a much lower calculation complexity, which is $4N_tN_r + 4N_t$. Compared with these methods of calculating Gram matrix first and then multiplying with vector, the complexity, in this case, is greatly reduced. Since K and L are much smaller than N_t and N_r for massive MIMO systems, the complexity of the EPA-SORI algorithm is much lower than that of other algorithms.

7. Simulation Result

In this section, the BER performance results of the EPA-SORI algorithm are presented and compared with EPA-wNSA [32], MMSE [10], EP-INSA [9], and the Exact EP [18] algorithms. To fully demonstrate and verify the BER performance of the proposed EPA-SORI, simulations are performed under different modulation methods (i.e., 16/64/256QAM) and different loading factors (i.e., $\zeta = 0.5/0.25$). For some damping factor $\vartheta \in [0, 1]$, we set $\vartheta = 0.5$ in EPA-wNSA and EPA-SPRI algorithms according to [20]. Assuming that the base station is able to obtain perfect channel state information (CSI) and complete signal detection based on the obtained CSI.

To further exhibit the convergence performance of EPA-SORI, Error-vector magnitude (EVM), which is defined as $EVM = (\|\hat{\mathbf{s}} - \mathbf{s}\|_2/\|\mathbf{s}\|_2) \times 100\%$ [33], is considered in Figure 2. As presented in Figure 2, EPA-INSA diverges for three different modulation methods when $N_r = 128$, $N_t = 64$. In contrast, the convergence performance of EPA-wNSA can be improved by the optimal choice of the weighted factor, but the degree of improvement is very limited [32]. At the same time, the proposed EPA-SORI algorithm could fast converge to the accurate Exact EP algorithm with obviously fewer iterations. As the modulation order increases, the benefit brought by EPA-SORI is more obvious. Therefore, it is verified that the advantage of the proposed algorithm is in fast convergence.

As shown in Figures 3-5, when the Massive MIMO system is configured as 32×64, 32×128, or 64×128, the loading factor ζ is 0.5 and 0.25. In each system configuration, we consider three modulation methods: 16QAM, 64QAM, and 256QAM, and different algorithms are compared and analyzed. In Figure 4, at BER 10^{-3} with $N_r = 128$, $N_t = 32$ for 256-QAM, EPA-SORI $(K = 3, L = 7)$ has a better performance than that of MMSE 0.9 dB. And in Figure 3, at BER 10^{-3} with $N_r = 64$, $N_t = 32$ for 256-QAM, EPA-SORI $(K = 9, L = 7)$ outperforms MMSE 2.2 dB. It can be seen that the BER of EPA-SORI algorithm is obviously superior to that of the MMSE, and as the loading factor ζ increases, the advantage will grow further. In Figures 4 and 5, EPA-INSA $(K = 6, L = 7)$ has a good performance when $\zeta = 0.25$, but does not converge when $\zeta = 0.5$. Meanwhile, choosing the weight factor of EPA-wNSA helps improve the convergence performance of EPA-wNSA but makes it difficult for further improvement, especially at a low ζ. In contrast, EPA-SORI uses SORI iteration to solve the only one-time matrix inversion in EPA, which enables it to fast converge to the accurate Exact EP algorithm with low complexity. For example, in Figure 4, at BER 10^{-3} with $N_r = 128, N_t = 32$ for 64-QAM, EPA-SORI $(K = 3, L = 7)$ outperforms EPA-wNSA $(K = 5, L = 7)$ 0.3 dB, and for 256-QAM, EPA-SORI $(K = 3, L = 7)$ outperforms EPA-wNSA $(K = 5, L = 7)$ 3.1 dB. In Figure 5, at BER 10^{-3} with $N_r = 128, N_t = 64$ for 64-QAM, with only 8 iterations used by EPA-SORI $(K = 8, L = 7)$, its performance is outperforming that of EPA-wNSA algorithm $(K = 21, L = 7)$ 2 dB which requires 21 iterations 0.2 dB. And for 256-QAM, the advantage will grow further. In other words, the performance of the EPA-SORI algorithm is always superior to that of EPA-wNSA under the same Massive MIMO system configuration.

Furthermore, a clear overview of the performance-complexity trade-off under different modulation methods and system configurations is provided in Figure 6. From Figure 6, the EPA-SORI algorithm can achieve not only significantly better performance than MMSE with a significantly lower computational complexity, but also almost the same performance as the Exact EP. For example, in Figures 6(b) and 6(c), at BER 10^{-3} with $N_r = 64$, $N_t = 32$ for 64-QAM and 256-QAM, EPA-SORI $(K = 9, L = 7)$ outperforms MMSE 2.4 dB and 3.8 dB, respectively, which is very close to that of Exact EP. However, EPA-SORI $(K = 9, L = 7)$ only consumes 67% and 63% of computational cost of MMSE and Exact EP, respectively. It is also observed from Figure 6, EPA-wNSA can improve the BER performance by increasing the number of iterations K at the cost of a substantial increase in computational complexity. In contrast, the EPA-SORI algorithm can achieve BER performance close to Exact EP with fewer iterations, and its complexity is much lower than the Exact EP and EPA-wNSA. For example, in Figure 6(b), at BER 10^{-3} with $N_r = 64$, $N_t = 32$ for 64-QAM, by increasing the number of iterations K, the performance of EPA-wNSA $(K = 29, L = 7)$ is increased by 0.9 dB compared with EPA-wNSA $(K = 16, L = 7)$, but the complexity is increased by 11% compared with EPA-wNSA $(K = 16, L = 7)$. However, EPA-

Table 1: Computational complexities comparison.

Detector scheme	Computational complexity
MMSE [10]	$4N_t(N_t^2 + N_tN_r + N_r)$
Exact EP [18]	$4N_t(N_t^2 + N_tN_r + N_r) + L(4N_t^2 + 8N_t) + 4N_t$
EPA-INSA [9]	$4N_t^2N_r + 4KN_t^2 + 4N_tN_r + L(4N_t^2 + 8N_t) + 4N_t$
EPA-wNSA [32]	$4N_t^2N_r + (4K + 4)N_t^2 + 4N_tN_r + L(4N_t^2 + 8N_t) + 6N_t$
Proposed EPA-SORI	$(K + 1)(8N_tN_r + 6N_t) + L(8N_tN_r + 8N_t) + 4N_t$

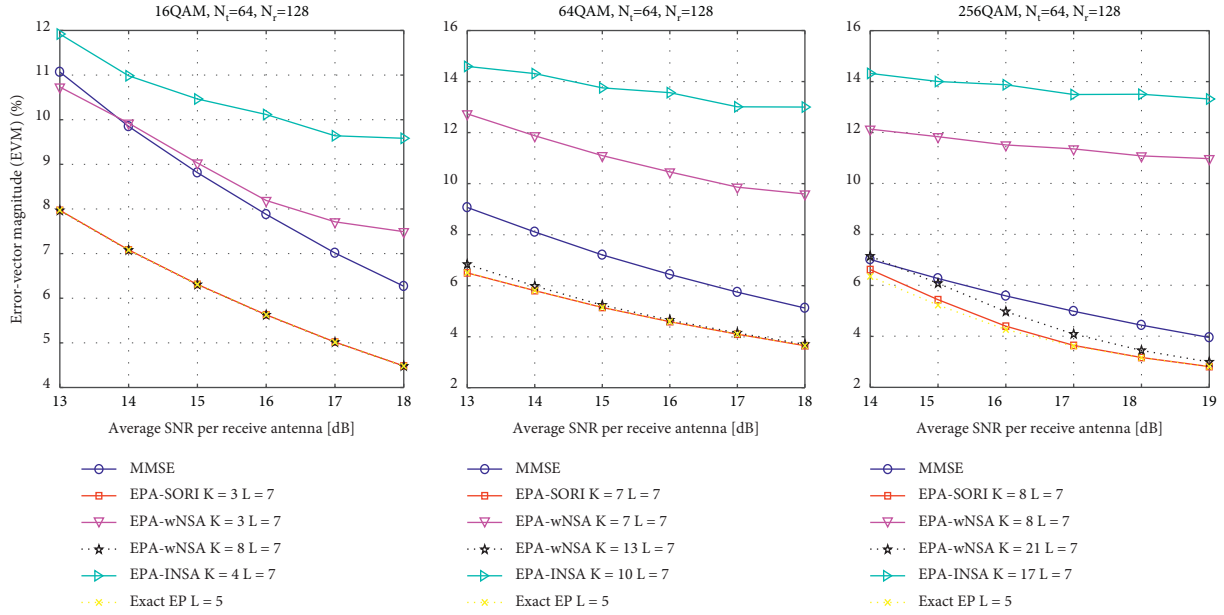

Figure 2: The error-vector magnitude comparison for different detectors under different modulations with fixed $N_t = 64$ and $N_r = 128$.

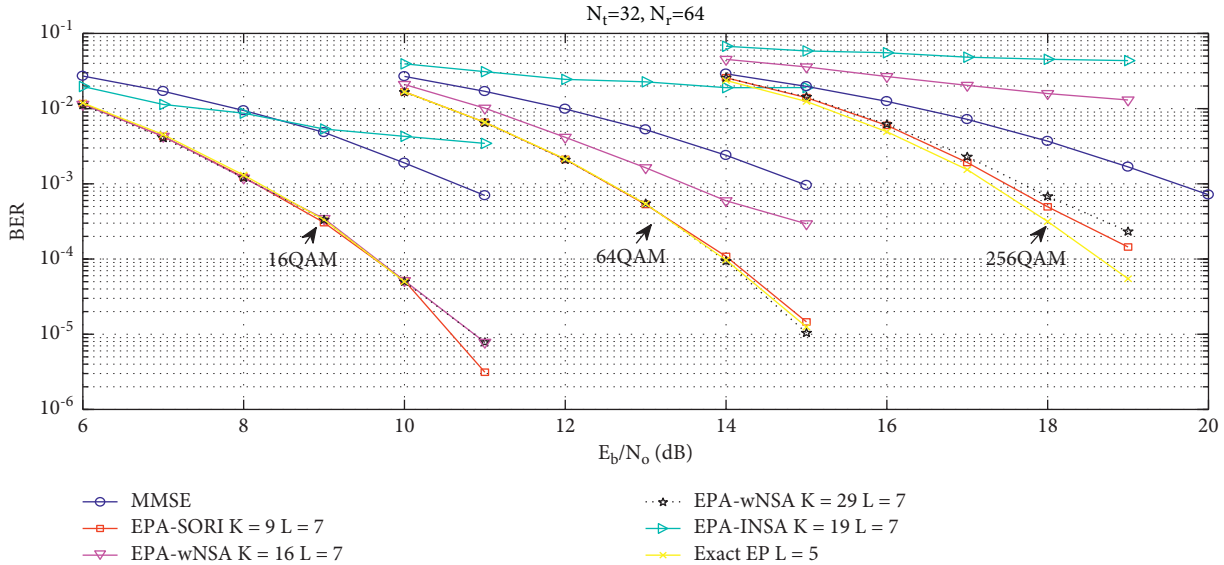

Figure 3: The BER performance comparison of the proposed EPA-SORI, MMSE, EPA-INSA, EPA-wNSA, and the Exact EP for different modulations. At the same time, the system configurations with $N_t = 32$ and $N_r = 64$ are considered.

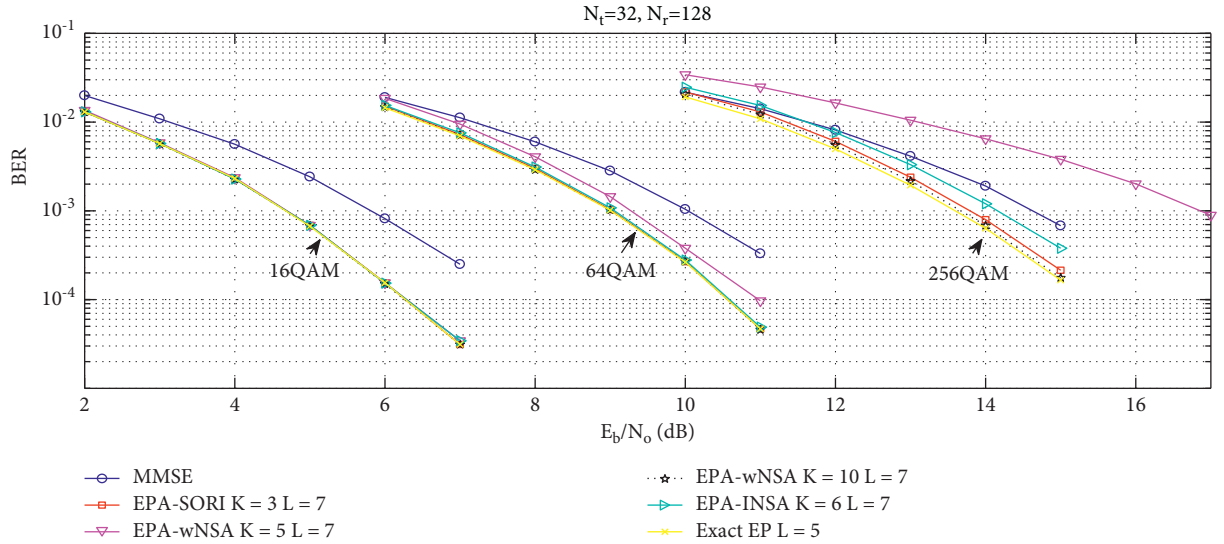

Figure 4: The BER performance comparison of the proposed EPA-SORI, MMSE, EPA-INSA, EPA-wNSA, and the Exact EP for different modulations, when $N_t = 32$ and $N_r = 128$.

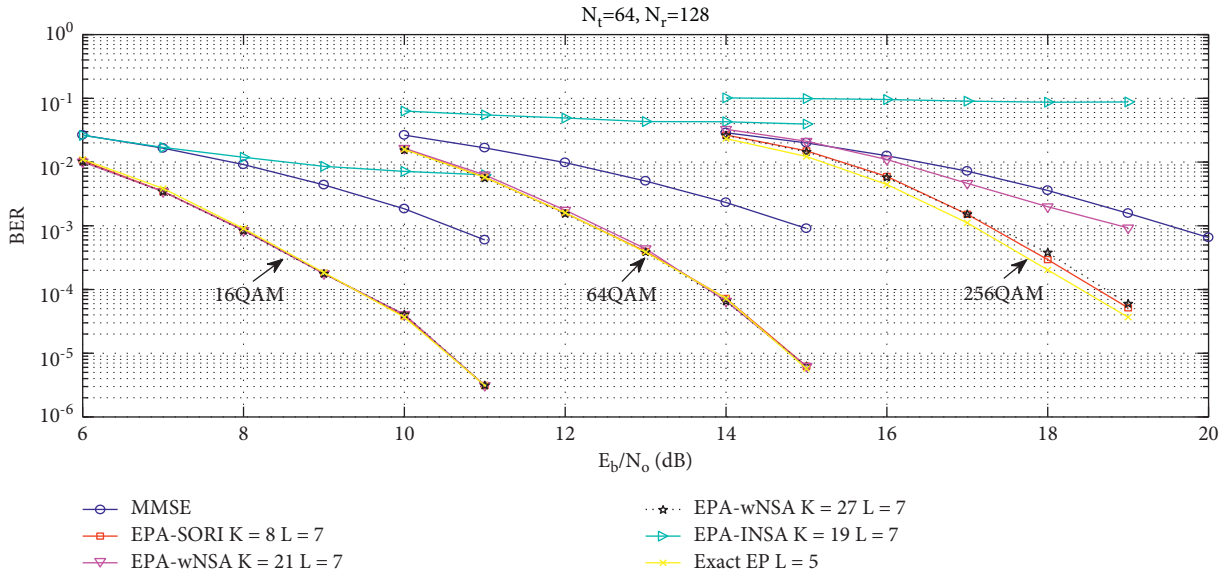

Figure 5: The BER performance comparison of the proposed EPA-SORI, MMSE, EPA-INSA, EPA-wNSA, and the Exact EP for different modulations when $N_t = 64$ and $N_r = 128$.

SORI $(K = 8, L = 7)$ can achieve a performance close to the Exact EP when $K = 8$, but it only consumes 66% of complexity of EPA-wNSA $(K = 29, L = 7)$. Next, we compare the EPA-SORI, EP-wNSA and Exact EP algorithms in Figures 6(a) and 6(d). In Figure 6(a), at BER 10^{-3} with $N_r = 128$, $N_t = 32$ for 256-QAM, EPA-SORI $(K = 3, L = 7)$ outperforms EPA-wNSA $(K = 5, L = 7)$ 2.3 dB, and the complexity is 65% of EP-wNSA $(K = 10, L = 7)$ and 52% of that of the Exact EP. In addition, in Figure 6(d), at BER 10^{-3} with $N_r = 128$, $N_t = 64$ for 256-QAM, EPA-SORI $(K = 8, L = 7)$ outperforms EPA-wNSA $(K = 21, L = 7)$ 2 dB. At the same time, it only consumes 42% of the computational cost of EP-wNSA $(K = 27, L = 7)$ and 29% of that of Exact EP. In summary, compared with MMSE, Exact EP and the recently reported EPA-wNSA algorithms, the proposed EPA-SORI algorithm has a better performance-complexity trade-off advantage,

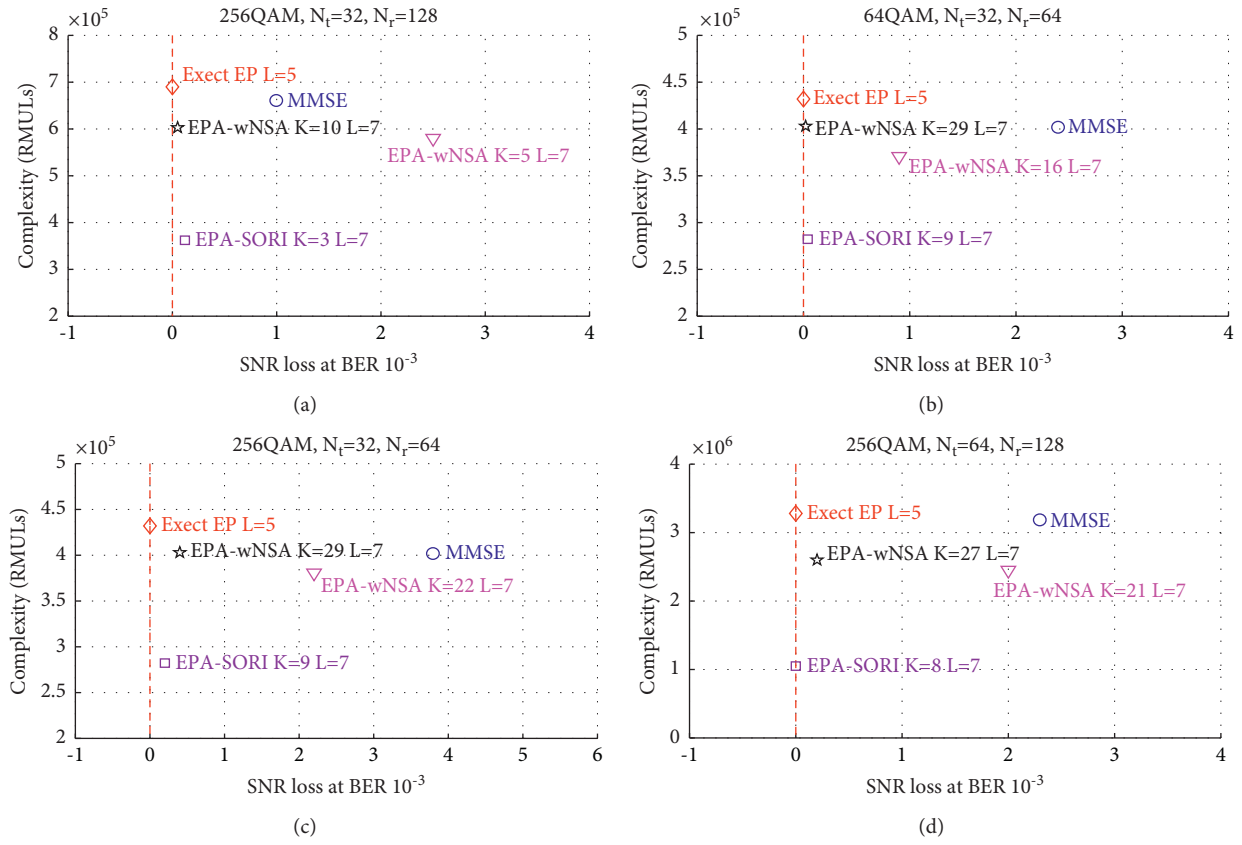

Figure 6: The performance-complexity trade-off comparison for different detectors under different system configurations. Exact EP is set as the performance benchmark to compare the SNR loss of MMSE, EPA-wNSA detectors. EPA-INSA detector is not compared in this figure because it fails to converge results in very poor performance, which cannot achieve the illustrated BER level.

which is more obvious in scenarios with high modulation order and a large number of users.

8. Conclusion

In this paper, we propose a novel data-detection scheme, EPA-SORI detector, which can achieve the same BER performance as the Exact EP algorithm with significantly lower complexity in various massive MIMO system configurations. At the same time, compared with MMSE and the existing EPA algorithms, the proposed EPA-SORI algorithm has a better performance-complexity trade-off advantage, which is more obvious in scenarios with high modulation order and a large number of users. The proposed algorithm avoids the direct calculation of the Gram matrix. At the same time, several effective techniques (i.e., the iteration initial solution and the optimal relaxation factor) are adopted to further enhance the convergence rate and accuracy.

In future work, there will be many potential applications. The proposed design can be extended to other more complex scenarios, such as the extension of EPA-SORI to decentralized architectures [34–36]. Also, it can be further combined with deep learning methods to improve performance [37, 38]. Finally, we will investigate the proposed design to more realistic channel scenarios in our future work.

Data Availability

The data used to support the findings of this study are included within the article. No other data were used beyond those in this article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Hainan Province under Grants 2019RC130 and 620QN238, in part by the National Natural Science Foundation of China under Grant 61771066, and in part by the Scientific Research Fund Project of Hainan University under Grants KYQD(ZR)-1999, KYQD(ZR)-21007, and KYQD(ZR)-21008.

References

[1] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, 2013.

[2] J. Hu, Y. Wu, R. Chen, F. Shu, and J. Wang, "Optimal detection of UAV's transmission with beam sweeping in covert

wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 1080–1085, 2020.

[3] J. Cespedes, P. M. Olmos, M. Sanchez-Fernandez, and F. Perez-Cruz, "Probabilistic MIMO symbol detection with expectation consistency approximate inference," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3481–3494, 2018.

[4] F. Rusek, D. Persson, B. K. Buon Kiong Lau et al., "Scaling up mimo: opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2013.

[5] X. Liang, W. Xu, H. Gao, M. Pan, and P. Zhang, "Throughput optimization for cognitive uav networks: a three-dimensional-location-aware approach," *IEEE Wireless Communications Letters*, vol. 9, no. 7, pp. 948–952, 2020.

[6] C. Li, F. Sun, J. M. Cioffi, and L. Yang, "Energy efficient mimo relay transmissions via joint power allocations," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 61, no. 7, pp. 531–535, 2014.

[7] C. Li, H. J. Yang, S. Fan, J. M. Cioffi, and L. Yang, "Multi-user overhearing for cooperative two-way multi-antenna relays," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 3796–3802, 2015.

[8] E. BjoRnson, J. Hoydis, and L. Sanguinetti, "Massive mimo networks: spectral, energy, and hardware efficiency," *Foundations and Trends in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[9] X. Tan, W. Xu, Y. Zhang, Z. Zhang, and C. Zhang, "Efficient expectation propagation massive mimo detector with neumann-series approximation," *IEEE Transactions Circuits and Systems II*, vol. 67, no. 10, pp. 1924–1928, 2020.

[10] M. A. Albreem, M. Juntti, and S. Shahabuddin, "Massive mimo detection techniques: a survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3109–3132, 2019.

[11] X. Zhou, Q. Wu, S. Yan, F. Shu, and J. Li, "Uav-enabled secure communications: joint trajectory and transmit power optimization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 4069–4073, 2019.

[12] X. Zhou, S. Yan, J. Hu, J. Sun, J. Li, and F. Shu, "Joint optimization of a uav's trajectory and transmit power for covert communications," *IEEE Transactions on Communications*, vol. 67, no. 16, pp. 4276–4290, 2018.

[13] A. Chawla, A. Patel, A. K. Jagannatham, and P. K. Varshney, "Distributed detection in massive mimo wireless sensor networks under perfect and imperfect csi," *IEEE Transactions on Signal Processing*, vol. 67, no. 15, pp. 4055–4068, 2019.

[14] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[15] J. Yang, W. Song, S. Zhang, X. You, and C. Zhang, "Low-complexity belief propagation detection for correlated large-scale mimo systems," *Journal of Signal Processing Systems*, vol. 90, no. 4, pp. 585–599, 2018.

[16] J. Yang, C. Zhang, L. Xiao, S. Xu, and X. You, "Improved symbol-based belief propagation detection for large-scale mimo," *Signal Processing Systems*, vol. 11, 2015.

[17] L. Kuang, D. D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser mimo-ofdm systems using approximate message passing," *IEEE Tournal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 902–915, 2014.

[18] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, "Expectation propagation detector for extra-large scale massive mimo," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2036–2051, 2020.

[19] I. Santos, J. J. Murillo-Fuentes, R. Boloix-Tortosa, E. Arias-De-Reyna, and P. M. Olmos, "Expectation propagation as turbo equalizer in isi channels," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 360–370, 2017.

[20] X. Tan, Y.-L. Ueng, Z. Zhang, X. You, and C. Zhang, "A low-complexity massive mimo detection based on approximate expectation propagation," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7260–7272, 2019.

[21] L. Dai, X. Gao, X. Su, and S. Han, I. Chih-Lin and Z. Wang, Low-complexity soft-output signal detection based on gauss-seidel method for uplink multi-user large-scale mimo systems," vol. 64, 2014.

[22] J. Zeng, J. Lin, and Z. Wang, "An improved gauss-seidel algorithm and its efficient architecture for massive mimo systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 9, pp. 1194–1198, 2018.

[23] C. Zhang, Z. Wu, C. Studer, Z. Zhang, and X. You, "Efficient soft-output gauss-seidel data detector for massive mimo systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 99, pp. 1–12, 2018.

[24] 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 *Proceedings of the 2014 IEEE Global Communications Conference*, pp. 3291–3295, Austin, TX, USA, December 2014.

[25] A. Yu, S. Jing, X. Tan et al., "Efficient successive over relaxation detectors for massive mimo," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 6, pp. 2128–2139, 2020.

[26] T. P. Minka, *Expectation Propagation for Approximate Bayesian Inference*, pp. 362–369, Morgan Kaufmann Publishers Inc., Burlington, MA, USA, 2013.

[27] M. Seeger, *Bayesian Gaussian Process Models: Pac-Bayesian Generalisation Error Bounds and Sparse Approximations*, University of Edinburgh, Scotland, U.K., 2003.

[28] J. Cespedes, P. M. Olmos, M. Sanchez-Fernandez, and F. Perez-Cruz, "Expectation propagation detection for high-order high-dimensional mimo systems," *IEEE Transactions on Communications*, vol. 62, no. 8, pp. 2840–2849, 2014.

[29] J. Tu, M. Lou, J. Jiang, D. Shu, and G. He, "An efficient massive mimo detector based on second-order richardson iteration: from algorithm to flexible architecture," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 11, pp. 4015–4028, 2020.

[30] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends™ in Communications and Information Theory*, vol. 1, no. 1, pp. 1–182, 2004.

[31] Q. Deng, L. Guo, C. Dong, J. Lin, D. Meng, and X. Chen, "High-throughput signal detection based on fast matrix inversion updates for uplink massive multiuser multiple-input multi-output systems," *IET Communications*, vol. 11, no. 14, pp. 2228–2235, 2017.

[32] X. Tan, H. Han, M. Li et al., "Approximate expectation propagation massive mimo detector with weighted neumann-series," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 2, pp. 662–666, 2021.

[33] H. A. Mahmoud and H. Arslan, "Error vector magnitude to snr conversion for nondata-aided receivers," *IEEE Transactions on Wireless Communications*, vol. 8, no. 5, pp. 2694–2704, 2009.

[34] K. Li, R. R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized baseband processing for massive mu-mimo systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 4, pp. 491–507, 2017.

[35] C. Jeon, K. Li, J. R. Cavallaro, and C. Studer, "Decentralized equalization with feedforward architectures for massive mu-mimo," *IEEE Transactions on Signal Processing*, vol. 67, no. 17, pp. 4418–4432, 2019.

[36] J. Rodriguez Sanchez, F. Rusek, O. Edfors, M. Sarajlic, and L. Liu, "Decentralized massive mimo processing exploring daisy-chain architecture and recursive algorithms," *IEEE Transactions on Signal Processing*, vol. 68, pp. 687–700, 2020.

[37] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: training deep neural networks for interference management," *IEEE Transactions on Signal Processing: A Publication of the IEEE Signal Processing Society*, vol. 66, no. 20, pp. 5438–5453, 2018.

[38] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2554–2564, 2019.