Efficient Matrix Polynomial Expansion Detector for Large-scale MIMO: An Inverse-Transform-Sampling Approach

Qian Deng, Xiaopeng Liang, *Member, IEEE,* Qiang Ni, *Senior Member, IEEE,* and Jinsong Wu, *Senior Member, IEEE*

***Abstract*—Matrix polynomial expansion (MPE) based detector incurs either complicated computation of polynomial coefficients or slow convergence in uplink large-scale MIMO (LS-MIMO) systems. To solve these issues, an improved MPE (IMPE) detector is proposed, which can speed up the convergence significantly with uncomplicated polynomial coefficients. However, a challenging problem of performing IMPE is needed to compute all the eigenvalues of channel covariance matrix in real time. Unfortunately, directly calculating the eigenvalues of the channel covariance matrix requires complexity, which is as costly as the matrix inverse. To this end, an inverse-transform-sampling based IMPE (ITS-IMPE) detector is proposed to enhance the convergence rate and accuracy in a simple way. First, the closed-form expression of the eigenvalue spectral cumulative distribution function of the channel covariance matrix is deduced analytically, which is a critical factor that influence the eigenvalues estimation. Second, the improved polynomial coefficients of ITS-IMPE are then introduced by a fast online ITS-based eigenvalues estimation algorithm and a least-squares fitting procedure, which achieve a well trade-off between precision and computation. Simulation results exhibit that ITS-IMPE detector is able to achieve a significant enhancement performance with much lower complexity compared with many reported detectors under Rayleigh fading channel and low spatial correlated channel.**

***Index Terms*—Large-scale MIMO, Inverse-transform-sampling, Matrix polynomial expansion, Fast convergence.**

I. Introduction

LARGE-SCALE multiple-input multiple-output (LS-MIMO) is regarded as a promising technology to enhance the spectral/energy efficiency requirements for the 5-th generation (5G) and beyond wireless systems [1], [2]. However, enormous system dimensions impose great challenges to signal detection in terms of practical implementations and computational complexity [3].

A. Literature Review on LS-MIMO Detector

The nonlinear maximum-likelihood (ML) detector is considered as the optimal detector, but its complexity increases exponentially in LS-MIMO, or high-order modulations that limit its applications in LTE, 5G and beyond [4], [5]. The variants of ML detectors like sphere decoding (SD) detector [6], K-best detector [7], reactive tabu search detector [8] and belief-propagation-based detector [9], exhibit near-optimal performance and lower complexity compared with ML detector. Nevertheless, these detectors are not the most suitable and popular choice for LS-MIMO circuit implementation. Other more complex detectors (e.g. approximate message passing (AMP) [10]) are able to provide superior performance, but they incur severe degradation for certain LS-MIMO configurations [11]. By comparison, approximate inversion-based linear detectors are popular for practical LS-MIMO systems, not only for their good performance-complexity trade-off, but also for easy implementation.

At present, a lot of attention has been paid to the approximate inversion-based linear detectors, among which the iterative methods based detectors (e.g., the variants of Gauss-Seidel (GS), the successive over relaxation (SSOR), and so on) [3], [12]–[15] and matrix polynomial expansion-based detectors [16], [17] exhibit advantages in performance and low-cost. Among them, the former detectors suffer from low parallelism and require calculation of the Gram matrix, while the latter ones have advantages in high parallelism and small computation latency [18]. It is widely known that one typical application scenario of 5G is ultra-reliable and low-latency communications (URLLC). Particularly, the end-to-end delay of URLLC service is maintained below 1ms [18], [19]. Therefore, the suitability of highly parallelizing detector algorithms is a critical determinant in meeting strict computational deadlines required for LS-MIMO in 5G and beyond wireless networks. But in this scenario, various matrix polynomial expansion-based detectors, such as Neumann Series Approximation (NSA) [20], [21], Cayley-Hamilton Theorem with maximum eigenvalue (CHTME) [22] and truncated polynomial expansion (TPE) [23], [24], suffer either slow convergence performance or complicated computation of polynomial coefficients in uplink LS-MIMO systems. In [23], [24], a TPE method based on Taylor series expansion was proposed to solve the precoding/detection problem. In this way, however, all polynomial coefficients need to be recalculated once the order of the polynomial changes, and the calculation of the corresponding polynomial coefficients was very complicated. To solve this problem, other simpler methods to calculate the asymptotic polynomial coefficients were proposed in [22],

Qian Deng, Xiaopeng Liang are with School of Information and Communication Engineering, Hainan University, Haikou 570228, China (e-mail: dqian108@163.com, liangxiaopeng315@163.com)(Corresponding author: Qian Deng).

Qiang Ni is with the School of Computing and Communications, InfoLab21, Lancaster University, Lancaster, LA1 4WA, U.K. e-mail: q.ni@lancaster.ac.uk.

Jinsong Wu is with the School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin 541004, China, and also with the Department of Electrical Engineering, Universidad de Chile, Santiago 9170124, Chile (e-mail: wujs@ieee.org).

[25]. Additionally, a weighted Neumann series approximation algorithm (wNSA) was proposed to approximate matrix inversion, in which only one parameter α was used to control matrix polynomial expansion coefficients a_l, $l \in \{0, 1, 2, \cdots, L-1\}$ (a_l is the l-th polynomial coefficient) [26]. Nevertheless, they incur a slow convergence performance.

B. Motivation and Contributions

In this paper, an improved matrix polynomial expansion (IMPE) detector is designed, which can speed up the convergence significantly with uncomplicated polynomial coefficients. Nevertheless, a challenging issue of designing these coefficients is needed to obtain all the eigenvalues of channel covariance matrix in real time. Unfortunately, the eigenvalues estimation can only be obtained offline through numerical integration and binary searching method [27], leading to unsatisfactory performance in practical LS-MIMO scenarios.

In fact, user terminals (such as mobile phones, unmanned aerial vehicles (UAVs) [28], sensors [29], vehicle to vehicle (V2V) [30], and so on) usually access and are removed from the cellular network randomly, which leads to the constant change of eigenvalues of the channel covariance matrix. Therefore, the fast estimation of all the eigenvalues online is necessary and critical.

To this end, based on the aforementioned IMPE, a novel inverse-transform-sampling based IMPE (ITS-IMPE) detector is further proposed. Unlike [23], the improved polynomial coefficients proposed herein are developed by a fast online ITS-based eigenvalues estimation algorithm and a least-squares fitting procedure, which achieve a well trade-off between precision and computation. Once the order of the matrix polynomial or the number of user changes, the ITS-IMPE algorithm will calculate the corresponding coefficients more simply and efficiently. Importantly, the ITS-IMPE algorithm are applicable to practical implementations not only for its close-to-MMSE performance, but also for its low complexity and hardware-friendly. The main contributions of this work are as follows:

1) We design an IMPE detector, which can speed up the convergence significantly with uncomplicated polynomial coefficients. In order to fast estimate all the eigenvalues online, we further proposed a novel ITS-IMPE detector, where all the eigenvalues estimation of the channel covariance matrix can be obtained at a very low computational complexity. It means that the negative impact on dimension change of practical LS-MIMO scenarios can be significantly reduced.
2) We derive a closed-form expression of the eigenvalue spectral cumulative distribution function (CDF) of the LS-MIMO channel covariance matrix, which is almost identical to the true CDF under different LS-MIMO system configurations. This is the key to implement ITS-IMPE detector. Furthermore, we deduce the improved polynomial coefficients by the ITS-based eigenvalue estimation algorithm and the least-squares fitting procedure, achieving a good trade-off between approximation accuracy and computation. When the order of the matrix polynomial changes, calculation of the polynomial coefficients becomes simple.
3) We verify the actual performance-complexity advantage of the proposed ITS-IMPE detector in LS-MIMO systems by simulation. In a word, the proposed ITS-IMPE exhibits favorable stability for different channel estimations and reaches a better trade-off between BER performance and complexity load under Rayleigh fading channel and low spatial correlation channel.

C. Outline of the Paper and Notations

Paper Outline: The remainder of this paper is organized as follows. The system model is introduced in Section II. IMPE and ITS-IMPE detectors are presented in Section III and Section IV, respectively. Next, the computational complexity is analyzed in Section V. And the simulation results are illustrated in Section VI. Finally, the conclusion is drawn in Section VII.

Notations: Matrices and column vectors are represented in uppercase and lowercase bold letters, respectively. $(\cdot)^*, (\cdot)^H, (\cdot)^{-1}, \Pr(\cdot)$, $\lambda(\cdot)$ and $\rho(\cdot)$ stand for the complex conjugate, conjugate transpose, inverse, probability, eigenvalue and spectral radius, respectively. $Un(a, b)$ denotes a uniform distribution with the lower endpoint a and upper endpoint b. Additionally, $diag(\mathbf{A})$ denotes the diagonal matrix, whose elements consist of diagonal elements of matrix $\mathbf{A}$. $[\mathbf{\Phi}]_{i,j}$ represents the (i, j)-th element in matrix $\mathbf{\Phi}$. $m \sim \mathcal{CN}(\mu, \xi)$ denotes that m is a complex random variable which obeys the Gaussian distribution with mean μ and variance ξ.

II. System Model

We consider an uplink multiuser LS-MIMO scenario with U single antenna users and B antennas at the base station side. The received signal can be expressed as: $\mathbf{y} = \mathbf{Hs} + \mathbf{n}$, where $\mathbf{H} \in {}^{B\times U}$ denotes the channel matrix, $\mathbf{n} = [n_1, \cdots, n_B] \in {}^{B\times 1}$ with $n_b \sim \mathcal{CN}(0, \sigma_n^2)$, $b \in \{1, \cdots, B\}$ denotes the additive white Gaussian noise vector, and $\mathbf{s}$ denotes a $(U \times 1)$-dimensional transmitted symbols from constellation set $\mathcal{A}$ with $\mathbb{E}\{\mathbf{ss}^H\} = \mathrm{E}_s\mathbf{I}_U$. At the base station side, the MMSE estimate of the transmitted symbols can be given by

$$\widehat{\mathbf{s}}_{MMSE} = (\mathbf{H}^H\mathbf{H} + \frac{\sigma_n^2}{E_s}\mathbf{I}_U)^{-1}\mathbf{H}^H\mathbf{y} = \mathbf{G}^{-1}\mathbf{H}^H\mathbf{y}, \tag{1}$$

where $\mathbf{G} = \mathbf{H}^H\mathbf{H} + \frac{\sigma_n^2}{E_s}\mathbf{I}_U$. Nevertheless, the direct calculation of the estimations in (1) incurs a high complexity of $\mathcal{O}(BU^2 + U^3)$ and unfavorable hardware implementation [22]. Furthermore, the computations are used for the Gram matrix $\mathbf{G}$ (i.e., $\mathcal{O}(BU^2)$), which are even higher than that of the matrix inverse especially when $B >> U$.

III. The Proposed IMPE detector

To alleviate the aforementioned issuse, a novel IMPE algorithm is proposed in this section, which does not require computation of the Gram matrix and can further enhance the convergence performance and accuracy of approximate inversion-based MPE algorithm.

According to the *Neumann series theorem* [11], [31], the invertible matrix $\mathbf{G}^{-1}$ can be expressed as: $\mathbf{G}^{-1} = \kappa \sum_{l=0}^{\infty} (\mathbf{I}_U - \kappa\mathbf{G})^l \approx \kappa \sum_{l=0}^{L-1} (\mathbf{I}_U - \kappa\mathbf{G})^l$ if $\lim_{l\to\infty} (\mathbf{I}_U - \kappa\mathbf{G})^l = 0$, which holds for $0 < \kappa < 2/\max\lambda(\mathbf{G})$ [23]. To speed up the convergence rate, a set of weighting factor $c_0, c_1, \cdots, c_{L-1}$ is introduced, thus $\mathbf{G}^{-1}$ is written as

$$\begin{aligned}\mathbf{G}^{-1} &= \lim_{\eta\to\infty} \kappa \textstyle\sum_{l=0}^{\eta} (\mathbf{I}_U - \kappa\mathbf{G})^l \\ &\approx \textstyle\sum_{l=0}^{L-1} c_l \kappa (\mathbf{I}_U - \kappa\mathbf{G})^l,\end{aligned} \tag{2}$$

where η and L denote the matrix polynomial terms. Let $\mathbf{G}_L \triangleq \sum_{l=0}^{L-1} c_l\kappa(\mathbf{I}_U - \kappa\mathbf{G})^l$. According to *binomial expansion theorem* [32], $\mathbf{G}^{-1}$ can be transformed into

$$\begin{aligned}\mathbf{G}^{-1} &\approx \mathbf{G}_L \\ &= \textstyle\sum_{l=0}^{L-1} \sum_{r=0}^{l} c_l \kappa \begin{pmatrix} l \\ r \end{pmatrix} (-\kappa)^l \mathbf{G}^l \\ &= a_0\mathbf{I}_U + a_1\mathbf{G} + a_2\mathbf{G}^2 + \cdots + a_{L-1}\mathbf{G}^{L-1}.\end{aligned} \tag{3}$$

Here, $\mathbf{G}^{-1}$ can be approximated by the first L-terms of a weighted matrix polynomial. Consider that the eigen-decomposition of $\mathbf{G} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, where $\mathbf{\Lambda} = diag\{\Lambda_1, \Lambda_2, \cdots, \Lambda_U\}$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_U]$. Here, $\Lambda_u (u = 1, 2, \cdots, U)$ is the u-th largest eigenvalue of $\mathbf{G}$, $\mathbf{v}_u \in \mathbb{C}^{U\times 1} (u = 1, 2, \cdots, U)$ is the u-th column eigenvector of $\mathbf{G}$. Substituting $\mathbf{G} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ into (3), we have

$$\begin{aligned}&\mathbf{V}\mathbf{\Lambda^{-1}}\mathbf{V}^{-1} \\ &\approx a_0\mathbf{V}\mathbf{V}^{-1} + a_1\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} + a_2\mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^{-1} + \cdots + a_{L-1}\mathbf{V}\mathbf{\Lambda}^{L-1}\mathbf{V}^{-1} \\ &= \mathbf{V}(\textstyle\sum_{l=0}^{L-1} a_l\mathbf{\Lambda}^l)\mathbf{V}^{-1},\end{aligned} \tag{4}$$

i.e., $\mathbf{\Lambda^{-1}} \approx \sum_{l=0}^{L-1} a_l\mathbf{\Lambda}^l$. Thus, without computing $\mathbf{V}$, $a_l (l = 1, 2, \cdots, L-1)$ can be obtained by solving the following linear equations

$$\begin{bmatrix} 1/\Lambda_1 \\ 1/\Lambda_2 \\ \vdots \\ 1/\Lambda_u \\ \vdots \\ 1/\Lambda_U \end{bmatrix} \approx \begin{bmatrix} 1 & \Lambda_1 & \cdots & \Lambda_1^{L-1} \\ 1 & \Lambda_2 & \cdots & \Lambda_2^{L-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \Lambda_u & \cdots & \Lambda_u^{L-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \Lambda_U & \cdots & \Lambda_U^{L-1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_u \\ \vdots \\ a_{L-1} \end{bmatrix}. \tag{5}$$

To solve linear equations (5), in this paper, we use the classic least-squares method with complexity $O(UL^2)$, where $L << U$. One of the most critical issue of solving (5) is to obtain eigenvalues $\Lambda_u (u = 1, 2, \cdots, U)$. The IMPE algorithm with the efficient polynomial coefficients $\{a_0, a_1, \cdots, a_l, \cdots, a_{L-1}\}$ is summarized in Algorithm 1.

Observe that the calculation of $\widehat{\mathbf{s}}_{MMSE} \approx (a_0\mathbf{I}_U + a_1\mathbf{G} + a_2\mathbf{G}^2 + \cdots + a_{L-1}\mathbf{G}^{L-1})\mathbf{H}^H\mathbf{y}$ involves multistage parallel multiplications of matrix-vector, such as $\mathbf{y}_3 \triangleq \mathbf{G}\mathbf{y}_1 = \mathbf{H}^H\mathbf{y}_2 + \frac{\sigma_n^2}{E_s}\mathbf{y}_1$ where $\mathbf{y}_2 \triangleq \mathbf{H}\mathbf{y}_1$ and $\mathbf{y}_1 \triangleq \mathbf{H}^H\mathbf{y}$. Nevertheless, directly calculating the eigenvalues of $\mathbf{G}$ requires the computational complexity of $O(U^3)$, which is as much as that of the matrix inverse. The work [27] proposed a an eigenvalue estimation approach from the specified distribution $f_\lambda(x)$, which is expressed as:

$$\begin{aligned}\int_{d_{l-1}}^{d_l} f_\lambda(x)dx &= (d_l - d_{l-1}) f_\lambda(x_u) \\ &= \tfrac{1}{U},\end{aligned} \tag{6}$$

Algorithm 1: The IMPE algorithm

Input: B, U, L, $\mathbf{y}$, $\mathbf{H} \in \mathbb{C}^{U\times B}$.
Output: $\mathbf{G}_L$.
1 Calculate the exact eigenvalues $\Lambda_u (u = 1, 2, \cdots, U)$ of $\mathbf{G}$.
2 $\mathbf{\Lambda} = diag\{\Lambda_1, \Lambda_2, \cdots, \Lambda_U\}$
3 Based on (5), the efficient polynomial coefficients $\{a_0, a_1, \cdots, a_l, \cdots, a_{L-1}\}$ are calculated.
4 Based on $\mathbf{V}\mathbf{\Lambda^{-1}}\mathbf{V}^{-1} = \mathbf{V}(\sum_{l=0}^{L-1} a_l\mathbf{\Lambda}^l)\mathbf{V}^{-1}$, $\mathbf{G}_L$ is calculated in (3).
5 $\widehat{\mathbf{s}} \approx \mathbf{G}_L\mathbf{H}^H\mathbf{y}$.

where $d_0 < d_1 < \cdots < d_U$ are the $U + 1$ points located on the horizontal axis. Here, the horizontal axis is x which represents the range of eigenvalues. And d_l is defined as: $\int_{d_0}^{d_l} f_\lambda(x)dx = l/U$, $l = 1, 2, \cdots, U-1$. After $\{d_l\}_{l=0}^{U}$ are calculated, the eigenvalues can be estimated as

$$x_u = f_\lambda^{-1}[\frac{1}{U(d_l - d_{l-1})}], \quad u = 1, 2, \cdots, U. \tag{7}$$

However, these points $d_1, \cdots, d_{U-1}$ can only be calculated offline through numerical integration and binary searching method [27], leading to unsatisfactory performance in practical scenarios. Furthermore, users randomly access and are removed from the cellular network, leading to the constant change of the number and the value of these points $d_1, \cdots, d_{U-1}$.

IV. The Proposed ITS-IMPE detector

To tackle the aforementioned issue, by making full use of the typical properties of LS-MIMO channel matrices, an ITS-IMPE detector is proposed in this section. Compared with other sampling methods, such as reject sampling and Monte Carlo methods, ITS method demonstrates excellent performance, greater simplicity and intuitiveness [33]. Next, we will detail the ITS-IMPE detector as follows.

First, a fast online ITS-based eigenvalues estimation algorithm consists of the following three steps:
1) *Step 1*: Deduce the closed-form expression of the CDF, e.g., $F_\lambda(x)$, of eigenvalues $\Lambda_u (u = 1, 2, \cdots, U)$ based on its empirical probability density distribution function;
2) *Step 2*: Calculate the inverse of the desired CDF which is obtained in *Step 1*, e.g., $F_\lambda^{-1}(x)$;
3) *Step 3*: Generate U uniform numbers within the interval $[0, 1]$ and substitute them into the inverse function of the CDF as input, then the results can be used as the eigenvalues estimation.

Second, the improved polynomial coefficients of ITS-IMPE is then introduced by the above on-line ITS-based eigenvalues estimation algorithm and a least-squares fitting procedure. The specific design and analysis is as follows.

A. Inverse-transform-sampling Scheme

Theorem 1. (Adapted from [33]): Let Z be the random variable, and its CDF is $F_Z \triangleq \Pr\{Z \le z\}, -\infty < z < +\infty$.

$F_Z^{-1}(m)$ is defined as $F_Z{}^{-1}(m) \triangleq \inf\{z : F_Z(z) = m\}$. If random variable $M \sim Un(0,1)$, then for all $z \in R$, the equation

$$\begin{aligned}\Pr(F_Z^{-1}(M) \le z) &= \Pr(\inf\{t : F_Z(t) = M\} \le z) \\ &= \Pr(M \le F_Z(z)) \\ &= F_M(F_Z(z)) \\ &= F_Z(z),\end{aligned} \tag{8}$$

where $\inf\{A\}$ and $Un(a,b)$ stand for the infimum of set A and a uniform distribution on the interval (a,b), respectively. Thus, Z satisfies $Z = F_Z^{-1}(M),\ \ M \sim Un(0,1)$.

Theorem 1 demonstrates that given a standard uniform distribution variable M on the interval $(0,1)$, and an invertible CDF F_Z, then the variable $Z = F_Z^{-1}(M)$ obeys the distribution of F_Z. That is to say, using **Theorem 1**, the number $M \sim Un(0,1)$ can be used to generate the sampling values, which are distributed as the specified distribution. However, the critical factor of the implementation of ITS scheme is to find the CDF, i.e, F_Z. An intuitive diagram of the processing of ITS scheme is presented in Fig. 1.

B. CDF of the Eigenvalues of LS-MIMO Channel Covariance Matrix

In the lemma below, we derive the closed-form expression of the eigenvalue empirical CDF of LS-MIMO channel covariance matrix for Rayleigh fading channels.

Lemma 1: In Rayleigh fading channels, for large-(B,U) regime with $U/B \to \tau$, the empirical cumulative distribution of channel covariance matrix eigenvalues $\lambda(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}})$ will almost surely converge to a certain limit distribution as

$$\begin{aligned}F_\lambda(x) &= \int_a^x f_\lambda(x)dx \\ &= \int_a^x (1-1/\tau)^+\delta(x) + \frac{\sqrt{(x-a)^+(b-x)^+}}{2\pi\tau x}dx \\ &= \frac{1}{2\pi\tau}[\sqrt{-x^2+2(1+\tau)x-(1-\tau)^2} \\ &\quad - |1-\tau|\arcsin\frac{3}{2\sqrt{\tau}x} + (1+\tau)\arcsin(\frac{1+\tau-x}{2\sqrt{\tau}}) + \pi\tau],\end{aligned} \tag{9}$$

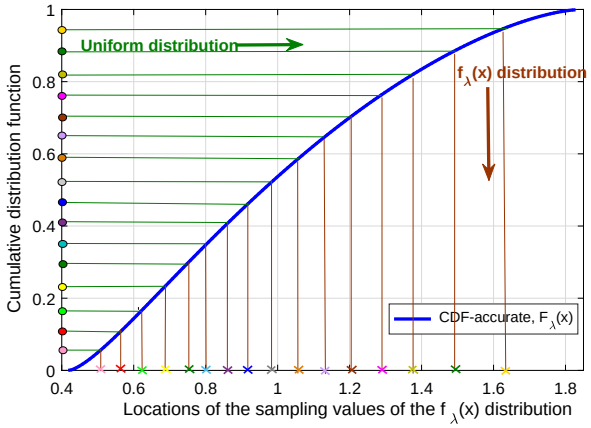

Fig. 1. The schematic diagram of ITS method.

where $\widehat{\mathbf{H}} = 1/\sqrt{B}\mathbf{H}$.

Proof: For LS-MIMO, with the Marcenko-Pastur distribution, empirical probability density function (PDF) of $\lambda(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}})$ converges to $f_\lambda(x)=(1-1/\tau)^+\delta(x) + \sqrt{(x-a)^+(b-x)^+}/2\pi\tau x$, where $(e)^+ = \max(0,e)$, $a = (1-\sqrt{\tau})^2$, and $b=(1+\sqrt{\tau})^2$ [34], [35]. Let $\Psi \triangleq (x-a)(b-x)$, $\theta \triangleq -ab$, $\vartheta \triangleq a+b$, $c=-1$, $\Delta \triangleq 4\theta c - \vartheta^2 = -(a-b)^2 < 0$. If $\theta < 0$, $\Delta < 0$, $c < 0$, $\sqrt{\vartheta^2 - 4\theta c} = (a+b)^2 - 4ab = b-a$, thus, we have

$$\frac{1}{2\pi\tau}\int\frac{\sqrt{\Psi}}{x}dx = \frac{1}{2\pi\tau}[\sqrt{\Psi} + \theta\int\frac{1}{x\sqrt{\Psi}}dx + \frac{\vartheta}{2}\int\frac{1}{\sqrt{\Psi}}dx], \tag{10}$$

where

$$\int\frac{1}{x\sqrt{\Psi}}dx = \frac{1}{\sqrt{-\theta}}\arcsin\frac{2\theta+\vartheta x}{x\sqrt{\vartheta^2-4\theta c}}, \tag{11}$$

$$\int\frac{1}{\sqrt{\Psi}}dx = -\arcsin(\frac{2cx+\vartheta}{\sqrt{-\Delta}}). \tag{12}$$

Substituting (11) and (12) into (10). Then we can obtain:

$$\begin{aligned}&\tfrac{1}{2\pi\tau}\int\tfrac{\sqrt{\Psi}}{x}dx \\ &= \tfrac{1}{2\pi\tau}[\sqrt{\Psi} - \sqrt{-\theta}\arcsin\tfrac{2\theta+\vartheta x}{x\sqrt{\vartheta^2-4\theta c}} - \tfrac{\vartheta}{2}\arcsin(\tfrac{2cx+\vartheta}{\sqrt{-\Delta}})] \\ &= \tfrac{1}{2\pi\tau}[\sqrt{\Psi} - \sqrt{ab}\arcsin\tfrac{(a+b)x-2ab}{x(b-a)} - \tfrac{a+b}{2}\arcsin(\tfrac{a+b-2x}{b-a})].\end{aligned} \tag{13}$$

Note that the range of $\lambda_u(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}})$, $u \in \{1,\cdots,U\}$ is $[a,b]$ [34]. According to the definition of CDF, we have

$$\begin{aligned}F_\lambda(x) &= \int_a^x f_\lambda(x)dx \\ &= \int_a^x (1-1/\tau)^+\delta(x) + \frac{\sqrt{(x-a)^+(b-x)^+}}{2\pi\tau x}dx \\ &= \tfrac{1}{2\pi\tau}[\sqrt{-x^2+(a+b)x-ab} - \sqrt{ab}\arcsin\tfrac{(a+b)x-2ab}{x(b-a)} \\ &\quad - \tfrac{(a+b)}{2}\arcsin(\tfrac{a+b-2x}{b-a}) - F_\lambda(a)],\end{aligned} \tag{14}$$

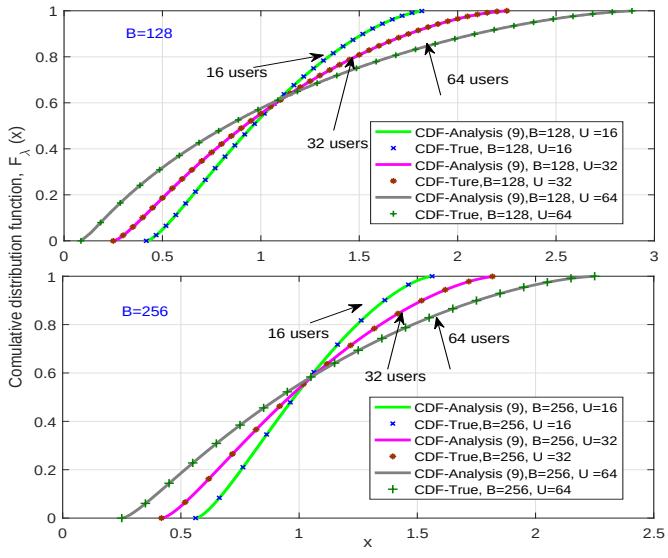

Fig. 2. CDF of the eigenvalues of $\widehat{\mathbf{H}}^H\widehat{\mathbf{H}}$ under different LS-MIMO system configurations.

$$F_\lambda(a) = \frac{1}{2\pi\tau}[\sqrt{-x^2+(a+b)x-ab} - \sqrt{ab}\arcsin\frac{(a+b)x-2ab}{x(b-a)} - \frac{(a+b)}{2}\arcsin(\frac{a+b-2x}{b-a})]\,|_{x=a} = \frac{1}{2\pi\tau}\frac{\pi(\sqrt{a}-\sqrt{b})^2}{4}. \quad (15)$$

Thus, a closed-form expression of $F_\lambda(x)$ can be obtained by

$$F_\lambda(x) = \frac{1}{2\pi\tau}[\sqrt{-x^2+(a+b)x-ab} - \sqrt{ab}\arcsin\frac{(a+b)x-2ab}{x(b-a)} - \frac{(a+b)}{2}\arcsin(\frac{a+b-2x}{b-a}) + \frac{\pi(\sqrt{a}-\sqrt{b})^2}{4}]. \quad (16)$$

Substituting $a = (1-\sqrt{\tau})^2$ and $b = (1+\sqrt{\tau})^2$ into (16). Thus, **Lemma 1** is proved. As presented in Fig. 2, the closed-form expression of CDF deduced by (9) is almost identical to that of the true CDF under different LS-MIMO system configurations. Therefore, it is a perfect match between our theoretical analysis and simulation results.

C. ITS-based Eigenvalues Estimation

Since $F_\lambda(x)$ is a monotonically increasing function (i.e., $m \le n \Rightarrow F_\lambda(m) \le F_\lambda(n)$), there exists an inverse function $F_\lambda^{-1}(x)$, $x \in [0,1)$, and $F_\lambda^{-1}(x)$ must also be monotonically increasing (i.e., $m \le n \Rightarrow F_\lambda^{-1}(m) \le F_\lambda^{-1}(n)$). According to (9), it is very difficult to calculate the analytical expression of $F_\lambda^{-1}(x)$. Thus, the linear approximation is used. More specifically, the interval $[a,b]$ is firstly partitioned into K segments, then we have $a = x_0 < x_1 < \cdots < x_k < \cdots < x_K = b$, $\forall k \in \{1,2,\cdots,K\}$. Each interval $\Delta_t = \frac{b-a}{K}$, so $x_k = x_0 + \Delta_t k$. Then, the linear function is calculated within $[a,b]$:

$$v' = F_\lambda(x_{k-1}) + \delta_k(x - x_{k-1}), \quad x_{k-1} \le x \le x_k, \quad (17)$$

where $\delta_k = \frac{F_\lambda(x_k)-F_\lambda(x_{k-1})}{x_k - x_{k-1}}$. Then the expression of the inverse of v' is

$$(v')^{-1} = \frac{x - F_\lambda(x_{k-1})}{\delta_k} + x_{k-1}, \quad F_\lambda(x_{k-1}) \le x \le F_\lambda(x_k). \quad (18)$$

Since $\lim_{K\to\infty}(v')^{-1} = F_\lambda^{-1}(x)$, thus, an approximately closed-form expression of the inverse function of CDF is given by

$$F_\lambda^{-1}(x) = \frac{x - F_\lambda(x_{k-1})}{\delta_k} + x_{k-1}, \quad F_\lambda(x_{k-1}) \le x \le F_\lambda(x_k). \quad (19)$$

Note that the value of K should be properly selected in practice to trade-off the approximate accuracy and complexity of the inverse function. Next, U uniform distributed samples $(x_{n1},\cdots,x_{nu},\cdots,x_{nU})$ are generated within the interval $(0,1)$. According to **Theorem 1**, we substitute $(x_{n1},\cdots,x_{nu},\cdots,x_{nU})$ into (19) as x and output the samples $F_\lambda^{-1}(x)$ which obey the PDF distribution. Thus, $F_\lambda^{-1}(x_{nu})$, $u \in \{1,2,\cdots U\}$ can be used as the estimated value of $\lambda_u(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}})$, $u \in \{1,2,\cdots,U\}$. Therefore, it yields the ITS-based eigenvalues estimation algorithm. Its implementation is detailed in Algorithm 2.

Since $\mathbf{G} = \mathbf{H}^H\mathbf{H} + \frac{\sigma_n^2}{E_s}\mathbf{I}_U$, the eigenvalues of $\mathbf{G}$ can be approximated by

$$\Lambda_u = B\lambda_u(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}}) + \frac{\sigma_n^2}{E_s} \approx BF_\lambda^{-1}(x_{nu}) + \frac{\sigma_n^2}{E_s}, \quad u \in \{1,2,\cdots,U\} \quad (20)$$

Algorithm 2: Online ITS-based eigenvalues estimation algorithm

Input: B, U.
Output: $\lambda_u(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}})$, $u \in \{1,2,\cdots,U\}$.
1 Compute $F_\lambda(x)$ by (9).
2 Compute U uniform distributed samples $(x_{n1},\cdots,x_{nu},\cdots,x_{nU})$ within the interval $(0,1)$.
3 Compute $\lambda_u(\widehat{\mathbf{H}}^H\widehat{\mathbf{H}})$, $u \in \{1,\cdots,U\}$ by (19).
4 Compute Λ_u, $u \in \{1,2,\cdots,U\}$ by (20).

As discussed above, once $F_\lambda(x)$ and $F_\lambda^{-1}(x)$ is deduced analytically, the eigenvalues estimation of $\mathbf{G}$ in the proposed ITS-based algorithm can be obtained with a very low complexity $O(U)$.

D. LLR Approximation

To enhance the accuracy of ITS-IMPE detector, the log-likehood ratios (LLRs) which can be extracted from soft decision [4], [36], is employed. Let $\mathbf{R} = \mathbf{G}^{-1}\mathbf{H}^H\mathbf{H}$ and $\mathbf{P} = \mathbf{R}\mathbf{G}^{-1}$, where the equivalent channel matrix is denoted by $\mathbf{R}$, and the equivalent channel gain is denoted by $\mu_u = \mathbf{R}_{(u,u)}$. Then, the noise-plus-interference (NPI) variance can expressed as

$$v_u^2 = \sum_{i\ne u}^{U} |\mathbf{R}_{(i,u)}|^2 + \mathbf{P}_{(i,i)}\sigma_n^2, \quad (21)$$

where $\mathbf{R}_{(i,u)}$ and $\mathbf{P}_{(i,i)}$ denote the (i,u)-th element in matrix $\mathbf{R}$ and the (i,i)-th element in matrix $\mathbf{P}$, respectively. The LLR of bit b for the u-th user can be computed by [4], [36]

$$L_{u,b} = \frac{\mu_u^2}{v_u^2}(\min_{q\in S_b^0}|\frac{\hat{s}_u}{\mu_u} - q|^2 - \min_{q'\in X_b^1}|\frac{\hat{s}_u}{\mu_u} - q'|^2). \quad (22)$$

Here, S_b^0 and S_b^1 are the b-th bit of the symbol from constellation set $\mathcal{A}$, where the b-th bit is 0 and 1. The proposed ITS-IMPE algorithm is summarized in Algorithm 3. Meanwhile, a functional diagram of ITS-IMPE processing and detection is given in Fig. 3.

E. Numerical Evaluation of Convergence Performance

We compare ITS-IMPE with wNSA [26] to demonstrate the convergence performance of ITS-IMPE detector. According to [26], the L–term wNSA is written as: $\mathbf{G}_L \triangleq \sum_{n=0}^{L-1}(\mathbf{I} - (\alpha\mathbf{D}^{-1})\mathbf{G})^n(\alpha\mathbf{D}^{-1})$, where $\mathbf{D} = diag(\mathbf{G})$, $0 < \alpha < 1$. By setting $\alpha = 1$, wNSA is called INSA [20]. Fig. 4 shows that, with fixed $B = 128$, the INSA diverges for $U = 32$ and $U = 40$. The reason is that when $B/U < 5.8$, the spectral radius of the convergence matrix of INSA $\rho(\mathbf{I}_U - \mathbf{D}^{-1}\mathbf{G}) > 1$ [37], and the INSA can not satisfy the condition for converging to $\mathbf{G}^{-1}$. At the same time, two preferable wNSA ($\alpha = 0.5$ (selected by [26]), $\alpha = 0.67$) converge at different rates for $U = 16$, $U = 24$, $U = 32$ and $U = 40$. By contrast, when the same L (e.g., $L > 2$) is used, the convergence performance of ITS-IMPE with different weighting factors $\{a_0, a_1, \cdots, a_l, \cdots, a_{L-1}\}$ is obviously enhanced compared with that of wNSA with a fixed

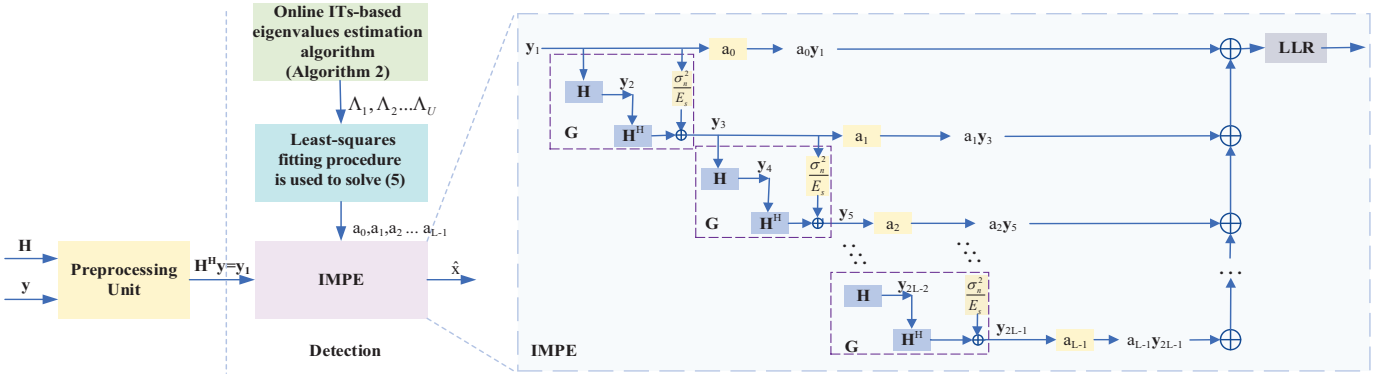

Fig. 3. Functional diagram of ITS-IMPE processing and detection.

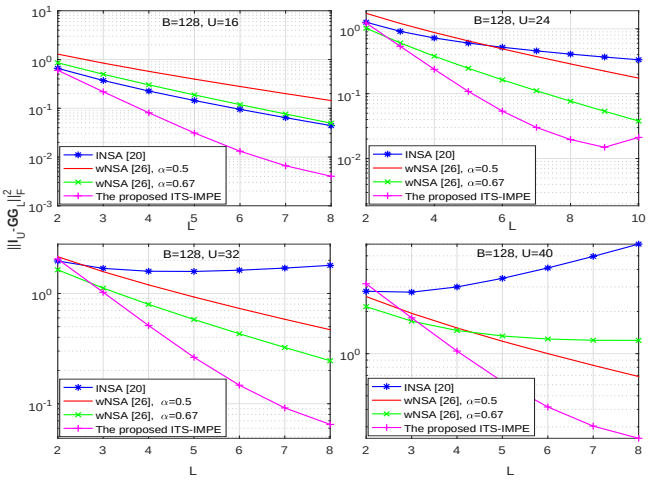

Fig. 4. The F-norms of $\|\mathbf{I}_U - \mathbf{G}\mathbf{G}_L\|_F^2$ with $B = 128$;

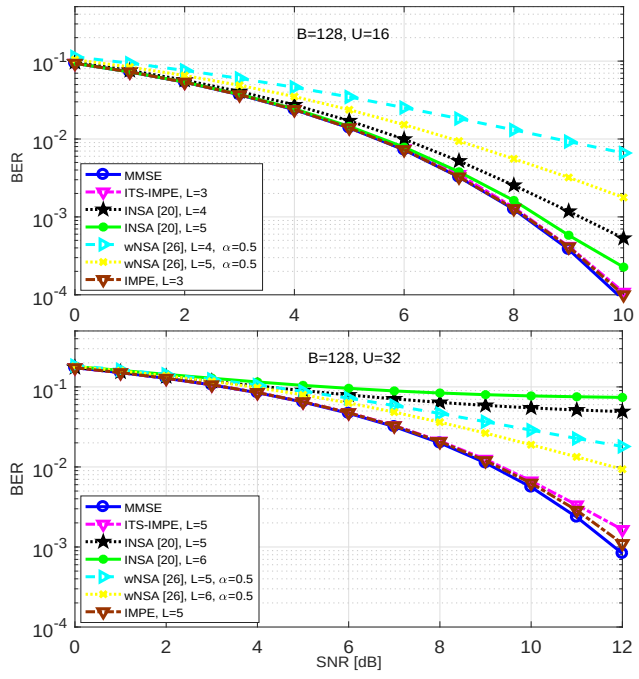

Fig. 5. The BER of ITS-IMPE, INSA and wNSA for $\alpha = 0.5$ with 16-QAM modulation.

Algorithm 3: The ITS-IMPE detector

Input: B, U, L, $\mathbf{y}$, $\mathbf{H} \in \mathbb{C}^{U\times B}$,$\mathbf{G} = \mathbf{H}^H\mathbf{H}$.
Output: $L_{u,b}$.

Generate U uniform numbers $(x_{n1}, \cdots, x_{nu}, \cdots, x_{nU})$ within the interval $[0, 1]$.
Compute τ, a, b by τ=U/B, $a = (1 - \sqrt{\tau})^2$, and b=$(1+\sqrt{\tau})^2$.
Based on PDF $f_\lambda(x)$, the empirical CDF of $\lambda_u(\widehat{\mathbf{H}}\widehat{\mathbf{H}}^H)$, $\forall u \in \{1, \cdots, U\}$ can be calculated by (9).
Based on CDF $F_\lambda(x)$, the approximately inverse function of the CDF $F_\lambda^{-1}(x)$ can be calculated by (19).
Plug $(x_{n1}, \cdots, x_{nu}, \cdots, x_{nU})$ into the approximately closed-form expression (19) as x, the estimated values of $\Lambda_u(u = 1, 2, \cdots, U)$ can be obtained by (20).
$\mathbf{\Lambda} = diag\{\Lambda_1, \Lambda_2, \cdots, \Lambda_U\}$
By using the least-squares fitting procedure to solve (5), the efficient polynomial coefficients $\{a_0, \cdots, a_l, \cdots, a_{L-1}\}$ are calculated.
Calculate $\widehat{\mathbf{s}} \approx (a_0\mathbf{I}_U + a_1\mathbf{G} + a_2\mathbf{G}^2 + \cdots + a_{L-1}\mathbf{G}^{L-1})\mathbf{H}^H\mathbf{y}$.
LLR Approximation:
for $u = 1 : U$ **do**
  for $b = 1 : \log_2|\mathcal{A}|$ **do**
    $L_{u,b} = \frac{\mu_u^2}{v_u^2}(\min_{q\in S_b^0}|\frac{\hat{s}_u}{\mu_u} - q|^2 - \min_{q'\in X_b^1}|\frac{\hat{s}_u}{\mu_u} - q'|^2)$.
  end
end

parameter. Furthermore, the benefit brought by the proposed ITS-IMPE can be shown in Fig. 5. As the number of users increased from 16 to 32, the bit error rate (BER) performance of INSA suffers obvious divergence since $\rho(\mathbf{I}_U - \mathbf{D}^{-1}\mathbf{G}) > 1$. Meanwhile, by choosing $\alpha = 0.5$ [26], the convergence performance of wNSA shows some improvement. Undoubtedly, the optimal choice of weight factor has a vital influence on the convergence performance. However, the improvement brought by wNSA is limited. This is due to the fact that only one parameter α is used to implement matrix polynomial expansion and the spectral radius of the iterative matrix $\rho(\mathbf{I}_U - (\alpha\mathbf{D}^{-1})\mathbf{G})$ increases with an increase in U. To further enhance the convergence rate and accuracy, by introducing efficient polynomial

coefficients $\{a_0, a_1, \cdots, a_l, \cdots, a_{L-1}\}$, the ITS-IMPE is able to achieve a significant enhancement performance. Furthermore, a performance bound of the ITS-IMPE algorithm which is obtained by IMPE algorithm is also used as the benchmark for comparison. Fig. 5 depicts that, without directly calculating the true eigenvalues and the complicated matrix inversion operations, ITS-IMPE algorithm can fast converge to the exact MMSE and its performance bound even when the number of users increases. This can be explained by the fact that the proposed ITS-based sample values can be used to estimate the true eigenvalues simply and effectively.

V. Complexity Analysis

Computational complexity analysis and comparison is given in this section.

For a given L, $\widehat{\mathbf{s}}$ can be obtained by

$$\begin{aligned}\widehat{\mathbf{s}} &\approx (a_0\mathbf{I}_U + a_1\mathbf{G} + a_2\mathbf{G}^2 + \cdots + a_{L-1}\mathbf{G}^{L-1})\mathbf{H}^H\mathbf{y}\\ &= a_0\mathbf{y}_1 + a_1\underbrace{(\mathbf{H}^H\mathbf{y}_2 + \frac{\sigma_n^2}{E_s}\mathbf{y}_1)}_{\mathbf{y}_3} + a_2\underbrace{(\mathbf{H}^H\mathbf{y}_4 + \frac{\sigma_n^2}{E_s}\mathbf{y}_3)}_{\mathbf{y}_5} + \cdots\\ &\quad + a_{L-1}\underbrace{(\mathbf{H}^H\mathbf{y}_{2L-2} + \frac{\sigma_n^2}{E_s}\mathbf{y}_{2L-3})}_{\mathbf{y}_{2L-1}}\\ &= a_0\mathbf{y}_1 + a_1\mathbf{y}_3 + a_2\mathbf{y}_5 + a_2\mathbf{y}_7 \cdots + a_{L-1}\mathbf{y}_{2L-1}\end{aligned} \quad , \quad (23)$$

$$\text{where} \begin{cases} \mathbf{y}_{2L-1} \triangleq \mathbf{G}\mathbf{y}_{2L-3} = \mathbf{H}^H\mathbf{y}_{2L-2} + \frac{\sigma_n^2}{E_s}\mathbf{y}_{2L-3} \\ \vdots \\ \mathbf{y}_3 \triangleq \mathbf{G}\mathbf{y}_1 = \mathbf{H}^H\mathbf{y}_2 + \frac{\sigma_n^2}{E_s}\mathbf{y}_1 \\ \mathbf{y}_2 \triangleq \mathbf{H}\mathbf{y}_1 \\ \mathbf{y}_1 \triangleq \mathbf{H}^H\mathbf{y} \end{cases} .$$

From (23), the calculation of $\widehat{\mathbf{s}}$ involves multistage parallel multiplications of matrix-vector (see Fig. 3 for detail). Hence, the complexity for the calculation of $\widehat{\mathbf{s}}$ is :

$$N = (2L-1)\,BU + 2UL - U. \quad (24)$$

Note that the number of complex-valued multiplications (NCMs) [38] is considered in the computation complexity. In addition, the number of NCMs involved in estimating the eigenvalues and performing least-squares method is approximately $N_{LS} = 3U + UL^2$ ($U >> L$), and the complexity of LLR approximation is: $N_{LLR} = U^2 + U$. Thus, without computing the Gram matrix, the overall complexity of ITS-IMPE can be calculated by:

$$\begin{aligned}N_{\text{total}} &= N + N_{LS} + N_{LLR}\\ &= (2L-1)\,BU + 2UL + 3U + UL^2 + U^2,\end{aligned} \quad (25)$$

which is evidently lower than many of the reported methods (see TABLE I). For the fair comparison of complexity and performance, the max-log LLR approximation is taken into account by all detectors in TABLE I. L_1 denotes the initial settings of Gauss-Seidel-NSA (GS-NSA) and RSI (i.e., the first L_1-terms of NSA in GS-NSA [39] and the first L_1 initial value of RSI [40]). L_{it} denotes the number of iterations in CG, QRI, CD, GS-NSA, JS-CD and RSI algorithms. Also note that Gram matrix calculations cannot be avoided in many of the reported methods, such as GS-NSA, JC-SD, CG, RSI and so on.

TABLE I
Computational complexities comparison

Detector scheme	Computational Complexity
MMSE	$((B+7)U^2 + U^3 + 3BU + 2U)/2$
NSA/wNSA($L \geq 2$) [20]	$(B+1)U^2 + BU + UL^2 + U^3 + U$
CG [41] [22]	$(BU^2 + 3BU)/2 + (2U^2 + 7U)(L_{it} - 1) + U + U^2$
QRI [42]	$(B/2+1)U^2 + 2BU + (2U^2 + U)L_{it} + U$
CD [43]	$L_{it}(4BU^2 + 2U) + U + U^2$
CHTME [22]	$(2L+2)BU + (3L-1)U + U^2$
JC-SD [44]	$2BU^2 + 4BU + 5U^2 + 13U + 4L_{it}U^2$
GS-NSA [39]	$BU^2/2 + (L_{it} + 2L_1 - 1)U^2 + 4U$
RSI [40]	$BU^2/2 + BU + 2U^2 + 4U + (U^2 + 2U + 4)L_{\text{it}}$
Proposed ITS-IMPE	$(2L-1)\,BU + 2UL + 3U + UL^2 + U^2$

VI. Simulation Results

To validate the performance of proposed ITS-IMPE detector, the BER of ITS-IMPE is studied and compared with exact MMSE, CG [41], QRI [42], CHTME [22], JCSD [44], GS-NSA [39], RSI [40] under perfect channel estimation and maximum likelihood (ML) channel estimation [45]. For simulations, independent and identically distributed (i.i.d.) Rayleigh fading channel, spatial correlation channel, and an uncoded LS-MIMO system with $B = 128$ and $U = \{16, 32, 40\}$ are considered. K is set to 8. At the same time, two modulations, 16-QAM and 64-QAM, are adopted.

First, the i.i.d. Rayleigh fading channel is taken into account. As presented in Fig. 6, using $L = 3$ is sufficient for ITS-IMPE to fast achieve almost the same BER performance as the exact MMSE, performing better over all the considering modulations under perfect channel estimation, when $B = 128, U = 16$. Furthermore, the proposed ITS-IMPE is superior to that of other reported algorithms such as CG, QRI, GS-NSA, CHTME, RSI and JCSD, which have apparently low complexity (see TABLE I and Fig. 8), especially when 64-QAM is employed. Moreover, with the decrease of base-station to user antenna ratios (BUARs), CHTME and JCSD suffer severe divergence. And at the same time, CG and QRI incur slow convergence performance, but ITS-IMPE can still obtain the similar near-optimal BER performance as exact MMSE. Thus, more user terminals can be served in the cellular network (such as sensors [29], unmanned aerial vehicles (UAVs) [28], vehicle to vehicle (V2V) [30], etc). As shown in Fig. 6, at BER 10^{-4} with $B = 128, U = 32$ for 16-QAM, ITS-IMPE ($L = 6$) outperforms the CG ($L_{it} = 6$) by 1.5 dB, outperforms GS-NSA ($L_1 = 2, L_{it} = 2$) by 2 dB, outperforms QRI ($L_{it} = 6$) by 2.1 dB and outperforms RSI ($L_1 = 2, L_{it} = 4$) by 0.5 dB, but at BER 10^{-4} with $B = 128, U = 40$ for 16-QAM, ITS-IMPE ($L = 7$) outperforms the CG ($L_{it} = 7$) by 4.9 dB, outperforms QRI ($L_{it} = 7$) by 5.2 dB. Additionally, GS-NSA ($L_1 = 2, L_{it} = 4$) and RSI ($L_1 = 2, L_{it} = 5$) with higher iterations not only is inferior to ITS-IMPE by a 0.2 dB and a 0.4 dB gap, but also suffers higher complexity burden. For higher-order modulations (e.g., 64-QAM), the BER performance advantage of ITS-IMPE is further increasing. In other words, the proposed ITS-IMPE detector is able to achieve a significant enhancement performance, and its complexity advantage is discussed in Fig. 8.

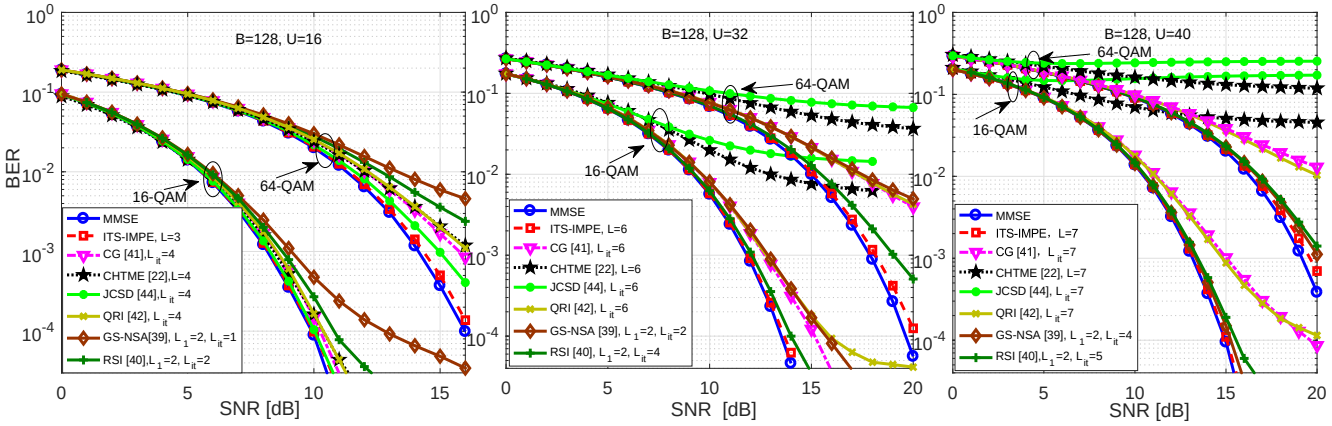

Fig. 6. BER performance comparison with perfect channel estimation in $B = 128$, $U = \{16, 32, 40\}$ LS-MIMO systems.

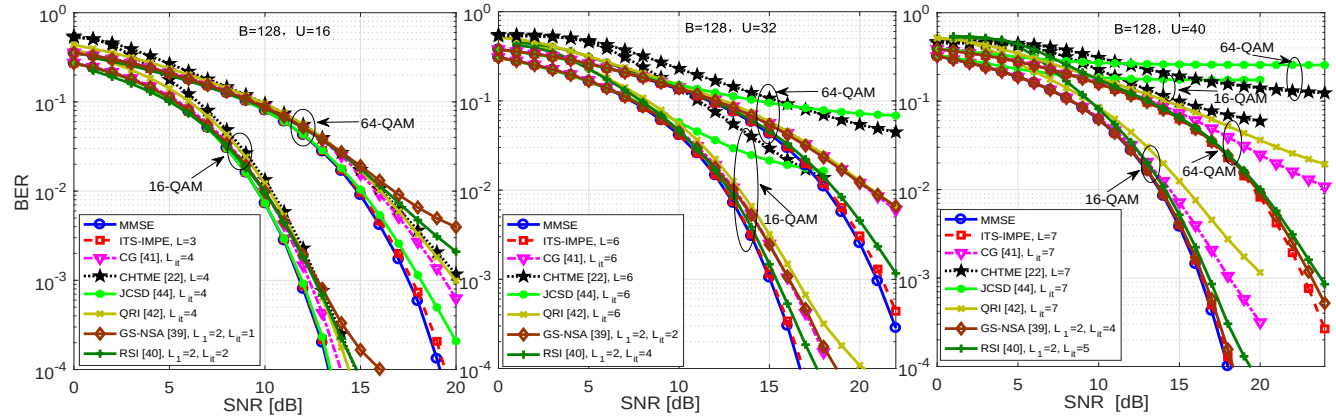

Fig. 7. BER performance comparison with ML channel estimator in $B = 128$, $U = \{16, 32, 40\}$ LS-MIMO systems.

Fig. 7 compares the BER performance of different algorithms with ML channel estimator under different LS-MIMO configurations. As observed in Fig. 7, the SNR required by exact MMSE to reach BER 10^{-4} increases compared with perfect channel estimation. Some detectors like CHTME and JCSD fail to converge when BUARs decreases (i.e., small B/U). However, the proposed ITS-IMPE not only maintains at a near-MMSE level, but also outperforms other algorithms even with smaller BUARs and higher-order modulations, when the ML channel estimator is employed. It further verifies the effectiveness and robustness of the ITS-IMPE detector.

Further, a clear overview of the performance-complexity trade-off under different channel estimations is provided in Fig. 8. As depicted in Fig. 8, the horizontal and vertical axes denote the SNR loss of each detection algorithm at a specific BER compared with exact MMSE and the complexities of the detection algorithms, respectively. On the horizontal axis, the smaller the SNR loss of the algorithms compared with exact MMSE, the better the algorithm performance. On the vertical axis, the less the NCMs, the lower the computational complexity. We find that the proposed ITS-IMPE detector, which comes with an improved matrix polynomial expansion to avoid matrix inversion operation, has a much lower computational cost compared with that of exact MMSE. At the

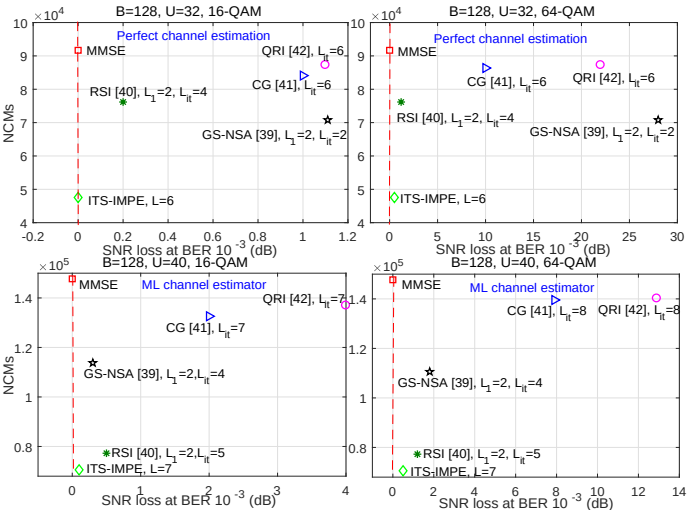

Fig. 8. The performance-complexity trade-off comparison for different detectors under different channel estimations. Exact MMSE is used as the benchmark to compare the SNR loss of ITS-IMPE, CG, GS-NSA, QRI detectors. JCSD and CHTME detectors is not compared in this figure, since the BER performance of these two detectors are declining sharply, which cannot achieve the illustrated BER level.

same time, its performance remains almost the same as the benchmark. From Fig. 8, with different channel estimations, ITS-IMPE is not only superior to GS-NSA, CG, QRI, but also consume obvious lower computational cost. For example, with the perfect channel estimation for 64-QAM, ITS-IMPE with 6 polynomial terms can achieve near MMSE performance at BER 10^{-3} when $B = 128, U = 32$, but the computational cost of ITS-IMPE is about 52.7% of MMSE. At the same time, the SNR loss of ITS-IMPE at BER 10^{-3} shows a decrease of 27.5 dB from that of GS-NSA ($L_1 = 2, L_{it} = 2$), 9.8 dB from that of CG ($L_{it} = 6$) and 22 dB from that of QRI ($L_{it} = 6$), respectively, but it only consumes 68%, 56.7% and 54% computational cost of that of GS-NSA, CG and QRI, respectively. In addition, ITS-IMPE with 6 terms is superior to RSI ($L_1 = 2, L_{it} = 4$) by a 0.2dB gap, but the computational cost of ITS-IMPE is about 64% of RSI. With ML channel estimation for 64-QAM, as shown in Fig. 8, the SNR loss of ITS-IMPE with 7 terms also shows a decrease of 1 dB from that of GS-NSA ($L_1 = 2, L_{it} = 4$), 7.5 dB from that of CG ($L_{it} = 8$), and 12.3 dB from that of QRI ($L_{it} = 8$), at BER 10^{-3} when $B = 128, U = 40$, but it only consumes 62.2%, 52% and 50% computational cost of GS-NSA, CG and QRI, respectively. In this case, RSI performs well, but the ITS-IMPE is still slightly better than it in both performance and complexity. Hence, the proposed ITS-IMPE exhibits favorable stability for different channel estimations and reaches a better trade-off between BER performance and complexity.

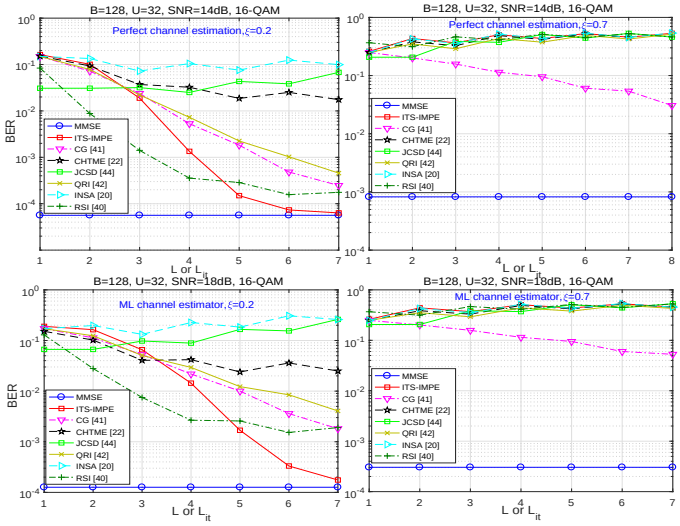

Fig. 9. Comparison of BER performance vs. L or L_{it} with correlation channel with 16-QAM modulation.

Finally, to further investigate the convergence performance of the different algorithms in correlation channel environment, the BER performance comparison among the detectors versus different L or L_{it} is given in Fig. 9. Note that the exponential correlation model which depicts the spatial correlation of LS-MIMO channels is adopted here [4], [23]. The exponential model is modeled as [46]: $[\mathbf{\Phi}]_{i,j} = \xi^{j-i}$, when $i \leq j$; $[\mathbf{\Phi}]_{i,j} = (\xi^{i-j})^*$, when $i > j$, where ξ $(0 \leq \xi \leq 1)$ represents magnitude of the correlation between two adjacent antennas at the base station side. Thus, spatially correlated channels, $\tilde{\mathbf{H}}$ is modeled as [47]: $\tilde{\mathbf{H}}=\mathbf{\Phi}^{1/2}\mathbf{H}$, where $\mathbf{H} \in \mathbb{C}^{B\times U}$ represents the Rayleigh fading coefficients with i.i.d. Gaussian elements. We made the BER performance of exact MMSE as the benchmark. From Fig. 9, INSA, CHTME and JCSD unable to converge, meanwhile CG and QRI suffer slow convergence performance in low correlation channel ($\xi = 0.2$) under both perfect and ML channel estimations, when $B = 128, U = 32$ for 16-QAM. As shown in Fig. 9, we can notice that RSI also fails to converge in the low correlation environment. This is due to the fact that RSI is a semi-iterative method based on Richardson iteration for signal detection, and it is difficult for Richardson iteration to produce more reliable iterative result for correlation channel. Additionally, as illustrated in Fig. 9, the proposed ITS-IMPE with a few numbers of polynomial terms can converge to exact MMSE under both perfect and ML channel estimations, which shows the performance advantage of ITS-IMPE in low correlated channel scenarios. For high correlated channel scenarios ($\xi = 0.7$), ITS-IMPE suffer severe divergence, it is because the empirical PDF is no longer applicable to high correlated channel environment [34], causing the deduced CDF (9) to be inapplicable. Therefore, obtaining the closed-form expression of eigenvalue CDF of channel covariance matrix with different propagation environments is the key to implement ITS-IMPE detector.

VII. Conclusion

Based on the proposed IMPE, we have further designed a novel data-detection scheme, ITS-IMPE detector, which can fast reach near-optimal performance of MMSE and its performance bound in both perfect channel estimation and ML channel estimation. The proposed detector does not require calculation of the Gram matrix and thus presents very low complexity. Furthermore, it is able to enable pipelining to execute multistage parallel structure for different modulations and hence exhibits low-cost and hardware-friendly. Theoretical and simulation results have shown the actual performance-complexity advantages of the proposed ITS-IMPE detector for LS-MIMO systems under Rayleigh fading channel and low spatial correlation channel.

As future work, there are many potential applications. The proposed design can be extended to other more complex detectors, such as expectation propagation with approximation (EPA) [26] [48] [49], ADMM-based infinity-norm (ADMIN) [36] [50], and so on, which involve underlying high-dimensional matrix inversion operations and outperform MMSE performance. Also, it can be used as a new iterative initial solution for other iterative algorithms, such as GS, CG, and so on. Furthermore, it can be used as a promising preconditioning matrix for some algorithms like Steepest Descent [37] to improve the convergence rate. Finally, the multilevel block matrix partition based on the correlation channel eigenvalue spectral CDF is further employed in the proposed design for high correlated channel scenarios, which is a part of ongoing work.

Acknowledgment

This work was supported in part by Hainan Provincial Natural Science Foundation of China under Grant 2019RC130

and Grant 620QN238, in part by the National Natural Science Foundation of China under Grant 61771066, and in part by the Scientific Research Starting Foundation of Hainan University under Grant KYQD(ZR)-1999 and Grant KYQD(ZR)-21132.

References

[1] X. You, C. Wang, J. Huang, X. Gao, and Z. Zhang, "Towards 6g wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Science China*, vol. 64, no. 1, pp. 1–76, 2021.

[2] B. C. Pandey, S. K. Mohammed, P. Raviteja, Y. Hong, and E. Viterbo, "Low complexity precoding and detection in multi-user massive mimo otfs downlink," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 5, pp. 4389–4405, 2021.

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

[4] L. Dai, X. Gao, X. Su, S. Han, . Chih-Lin, I., and Z. Wang, "Low-complexity soft-output signal detection based on gauss-seidel method for uplink multi-user large-scale mimo systems," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4839–4845, 2015.

[5] J. Fuentes, I. Santos, J. C. Aradillas, and M. Sanchez-Fernandez, "A low-complexity double ep-based detector for iterative detection and decoding in mimo," *IEEE Transactions on Communications*, vol. 69, no. 3, pp. 1538–1547, 2021.

[6] X. Chu and J. Mcallister, "Software-defined sphere decoding for fpga-based mimo detection," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 6017–6026, 2012.

[7] Tsai, P.-Y., Huang, and Z.-Y., "Efficient implementation of qr decomposition for gigabit mimo-ofdm systems," *IEEE Transactions on Circuits and Systems I Regular Papers*, vol. 58, no. 10, pp. 2531–2542, 2011.

[8] N. Srinidhi, S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "Low-complexity near-ml decoding of large non-orthogonal stbcs using reactive tabu search," in *2009 IEEE International Symposium on Information Theory*, 2009.

[9] 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.

[10] S. Wu, L. Kuang, Z. Ni, J. Lu, D. D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser mimo-ofdm systems using approximate message passing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 902–915, 2017.

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

[12] F. Jiang, C. Li, and Z. Gong, "Low complexity and fast processing algorithms for V2I massive MIMO uplink detection," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 6, pp. 5054–5068, Feb. 2018.

[13] 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. 68, no. 12, pp. 5049–5060, 2021.

[14] X. Gao, L. Dai, Y. Hu, Y. Zhang, and Z. Wang, "Low-complexity signal detection for large-scale mimo in optical wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 9, pp. 1903–1912, 2015.

[15] I. A. Khoso, X. Zhang, I. A. Khoso, Z. A. Dayo, and X. Dai, "Computationally efficient data detection in massive mimo wireless systems via semi-iterative method," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 252–10 264, 2021.

[16] H. Sifaou, A. Kammoun, L. Sanguinetti, M. Debbah, and M. S. Alouini, "Max-min sinr in large-scale single-cell mu-mimo: Asymptotic analysis and low-complexity transceivers," *IEEE Transactions on Signal Processing*, vol. 65, no. 7, pp. 1841–1854, 2017.

[17] A. A. Lu, X. Gao, Y. R. Zheng, and C. Xiao, "Low complexity polynomial expansion detector with deterministic equivalents of the moments of channel gram matrix for massive mimo uplink," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 586–600, 2016.

[18] A. Benzin, G. Caire, Y. Shadmi, and A. M. Tulino, "Low-complexity truncated polynomial expansion dl precoders and ul receivers for massive mimo in correlated channels," *IEEE Transactions on Wireless Communications*, vol. PP, no. 2, pp. 1069–1084, 2019.

[19] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, and H. Taoka, "Scenarios for 5g mobile and wireless communications: the vision of the metis project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, 2014.

[20] M. Wu, B. Yin, and G. Wang, "Large-scale MIMO detection for 3GPP LTE: algorithms and FPGA implementations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 916–929, Mar. 2014.

[21] M. Wu, B. Yin, K. Li, C. Dick, J. R. Cavallaro, and C. Studer, "Implicit vs. explicit approximate matrix inversion for wideband massive mu-mimo data detection," *Journal of Signal Processing Systems*, 2017.

[22] S. Shafivulla, A. Patel, and M. Z. A. Khan, "Low complexity signal detection for massive-mimo systems," *IEEE Wireless Communication Letters*, vol. 9, no. 9, pp. 1467–1470, 2020.

[23] A. Mueller, A. Kammoun, and E. BjoRnson, "Linear precoding based on polynomial expansion: reducing complexity in massive mimo," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, p. 63, 2016.

[24] A. Kammoun, A. Muller, E. Bjornson, and M. Debbah, "Linear precoding based on polynomial expansion: Large-scale multi-cell mimo systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 861–875, 2017.

[25] Q. Deng et al., "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.

[26] X. Tan, H. Han, M. Li, K. Sun, and C. Zhang, "Approximate expectation propagation massive mimo detector with weighted neumann-series," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 99, pp. 662–666, 2021.

[27] K. Lee and C. E. Chen, "An eigen-based approach for enhancing matrix inversion approximation in massive mimo systems," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5480–5484, 2017.

[28] 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.

[29] 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.

[30] 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.

[31] H. Wang, Y. Ji, Y. Shen, W. Song, and C. Zhang, "An efficient detector for massive mimo based on improved matrix partition," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2971–2986, 2021.

[32] B. A. Mamedov, "Analytical evaluation of the fully relativistic plasma dispersion function using binomial expansion theorems," *IEEE Transactions on Plasma Science*, vol. 37, no. 9, pp. 1770–1773, 2009.

[33] S. I. Gass and M. C. Fu, Eds., *Inverse Transform Method*. Boston, MA: Springer US, 2013, pp. 815–815. [Online]. Available: https://doi.org/10.1007/978-1-4419-1153-7˙200343

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

[35] G. Lu, J. Wu, R. C. Qiu, and H. Ling, "Analysis on the empirical spectral distribution of large sample covariance matrix and applications for large antenna array processing," *IEEE Access*, vol. 7, no. 1, pp. 2169–3536, 2019.

[36] S. Shahabuddin, I. Hautala, M. Juntti, and C. Studer, "Admm-based infinity-norm detection for massive mimo: Algorithm and vlsi architecture," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 29, no. 4, pp. 747–759, 2021.

[37] Q. Deng, X. Liang, X. Wang, M. Huang, and Y. Zhang, "Fast converging iterative precoding for massive mimo systems: An accelerated weighted neumann series-steepest descent approach," *IEEE Access*, vol. 8, pp. 50 244–50 255, 2020.

[38] W. Hui, J. Ya, and S. Yi, "An efficient detector for massive mimo based on improved matrix partition," *IEEE Transaction on Signal Processing*, vol. Early Access, no. 2, pp. 1–15, 2021.

[39] 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. 68, no. 12, pp. 5049–5060, 2021.

[40] I. Khoso, X. Zhang, I. Khoso, Z. Dayo, and X. Dai, "Computationally efficient data detection in massive mimo wireless systems via semi-

iterative method," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 252–10 264, 2021.
[41] Y. Hu, Z. Wang, X. Gao, and J. Ning, "Low-complexity signal detection using cg method for uplink large-scale mimo systems," in *IEEE International Conference on Communication Systems*, 2014.
[42] J. Minango and C. De Almeida, "Optimum and quasi-optimum relaxation parameters for low-complexity massive mimo detector based on richardson method," *Electron. Lett.*, vol. 53, no. 16, pp. 1114–1115, 2017.
[43] M. Wu, C. Dick, J. R. Cavallaro, and C. Studer, "High-throughput data detection for massive mu-mimo-ofdm using coordinate descent," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 12, pp. 2357–2367, 2017.
[44] X. Qin, Z. Yan, and G. He, "A near-optimal detection scheme based on joint steepest descent and jacobi method for uplink massive mimo systems," *IEEE Communications Letters*, vol. 20, no. 2, pp. 276–279, 2016.
[45] C. Yao, S. Zhang, and C. Pei, "Ml and map channel estimation for distributed one-way relay networks with orthogonal training," *China communications*, vol. 12, no. 12, pp. 84–91, 2015.
[46] H. Lim, Y. Jang, and D. Yoon, "Bounds for eigenvalues of spatial correlation matrices with the exponential model in mimo systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 11 961–204, 2017.
[47] B. M. Nouri, S. Haghighatshoar, and G. Caire, "Low-complexity statistically robust precoder/detector computation for massive mimo systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6516–6530, 2018.
[48] J. Cespedes, P. Olmos, M. P. 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.
[49] Y. Dong, C. Gong, Z. Zhang, X. Wang, K. Long, and X. Dai, "Low-complexity ep receiver based on location-aware and reliability-aware strategy," *IEEE Communications Letters*, vol. 25, no. 6, pp. 2034–2038, 2021.
[50] I. N. Tiba, Q. Zhang, J. Jiang, and Y. Wang, "A low-complexity admm-based massive mimo detectors via deep neural networks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4930–4934.

Qian Deng received the B.S. and M.S. degrees in Communication Engineering from Guilin University of Electronic Technology, China, in 2006 and 2009 respectively, and the Ph.D. degree with the Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2019. Since 2019, she is a lecturer with the College of Information and Communication Engineering, Hainan University. Her research interests include low-complexity precoding and detection in massive MIMO systems, cognitive wireless network, convex and nonconvex optimization with applications on signal processing, and UAV network.

Xiaopeng Liang received the B.S. degree in Communication Engineering from Heilongjiang University, China in 2006, and the M.S. degree in Communication Engineering from Guilin University of Electronic Technology, China in 2009. From Sept. 2016, he started to pursue the Ph.D. degree in Beijing University of Posts and Telecommunications (BUPT), China. Since 2021, he is a lecturer with the College of Information and Communication Engineering. His research interests focus on the massive MIMO signal processing, cognitive radio network and UAV network.

Qiang Ni (M04-SM08) received the B.Sc., M.Sc., and Ph.D. degrees from the Huazhong University of Science and Technology, China, all in engineering. He is currently a Professor and the Head of the Communication Systems Group, School of Computing and Communications, Lancaster University, Lancaster, U.K. His research interests include the area of future generation communications and networking, including green communications and networking, millimeter-wave wireless communications, cognitive radio network systems, non-orthogonal multiple access (NOMA), heterogeneous networks, 5G and 6G, SDN, cloud networks, energy harvesting, wireless information and power transfer, IoTs, cyber physical systems, AI and machine learning, big data analytics, and vehicular networks. He has authored or co-authored 300+ papers in these areas. He was an IEEE 802.11 Wireless Standard Working Group Voting Member and a contributor to various IEEE wireless standards.

Jinsong Wu (SM11) received PhD from Department of Electrical and Computer Engineering Queens University, Canada. He received 2020 IEEE Green Communications and Computing Technical Committee Distinguished Technical Achievement Recognition Award, for his outstanding technical leadership and achievement in green wireless communications and networking, He is elected Vice-Chair Technical Activities (2017-present), IEEE Environmental Engineering Initiative, a pan-IEEE effort under IEEE Technical Activities Board (TAB). He was the Founder and Founding Chair (2011-2017) of IEEE Technical Committee on Green Communications and Computing (TCGCC). He is also the co-founder and founding Vice-Chair (2015-present) of IEEE Technical Committee on Big Data (TCBD). He received 2017, 2019, and 2021 IEEE System Journal Best Paper Awards. His co-authored paper won 2018 IEEE TCGCC Best Magazine Paper Award. He was a Series Editor in the IEEE JOURNAL OF SELECTED AREAS ON COMMUNICATIONS Series on Green Communications and Networking. He is the Founder and the Editor of the IEEE Series on Green Communication and Computing Networks in the IEEE Communications Magazine. He is an Area Editor of the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING.