Efficient Statistical Linear Precoding for Downlink Massive MIMO Systems

Zheng Wang, *Senior Member, IEEE*, Le Liang, *Member, IEEE*, Shanxiang Lyu, Yili Xia, *Member, IEEE*, Yongming Huang, *Senior Member, IEEE*, and Derrick Wing Kwan Ng, *Fellow, IEEE*

***Abstract*— In this paper, we study low-complexity linear precoding for downlink massive multiple-input multiple-output (MIMO) systems, exploiting a statistical method. In sharp contrast to traditional linear precoding algorithms, our proposed efficient randomized iterative precoding algorithm (ERIPA) not only avoids costly matrix inversion but also considers the complexity reduction of matrix multiplication involved, thus enabling more efficient linear precoding. Additionally, ERIPA is demonstrated to have both exponentially fast and global convergence, making it adaptable to various practical scenarios of massive MIMO. We also investigate the convergence phenomenon of ERIPA in relation to the selection of the sampling distribution during random iterations. After that, the concept of conditional sampling is introduced to ERIPA such that significant system potential can be beneficially exploited in terms of both precoding performance and computational complexity. Finally, simulation results regarding the downlink massive MIMO are presented to confirm the superiorities of the proposed ERIPA.**

***Index Terms*— Massive MIMO, linear precoding, low complexity, global convergence, iterative methods, convergence analysis and enhancement.**

I. Introduction

MASSIVE multiple-input multiple-output (MIMO), which is one of the key technologies for enabling beyond fifth-generation (B5G) and sixth-generation (6G) wireless communications, is capable of dramatically improving the spectral efficiency and power efficiency tradeoffs [1], [2], [3], [4]. However, the downlink performance of massive MIMO is severely limited by the existence of multiuser interference. As a remedy, the technique of precoding is widely accepted for mitigating multi-access interference [5], [6], [7], [8], [9]. By targeting the transmitted signals at the intended receivers, various linear precoding schemes, such as zero-forcing (ZF) and regularized ZF (RZF), can approach the channel capacity when the number of antennas (denoted by N) at the base station (BS) is sufficiently higher than that employed at the user side (denoted by K), i.e., $N \gg K$ [10], [11], [12]. Unfortunately, both ZF and RZF involve complicated matrix computations such as matrix inversion and matrix multiplication. For example, the complexity order of matrix inversion is $O(K^3)$ that hinders the practical implementation of ZF or RZF, especially for massive MIMO systems with a high-dimensional downlink. Towards this end, various low-complexity linear precoding schemes have been designed to circumvent the need for matrix inversion [13], [14], [15], [16], [17], [18], [19].

In particular, in [20], the Jacobi iterative method is applied for low-complexity linear precoding. However, the Jacobi iteration-based precoding only works when $N \gg K$. In fact, besides the Jacobi iterative method, other low-complexity precoding schemes based on the series expansions like Neumann series (NS), Newton iteration (NI), also have to fulfill the same convergence requirement [21], [22], [23], [24], [25], making them impractical for numerous cases of interest. On the other hand, in [26], the Richardson iteration is employed for linear precoding with an introduced relaxation coefficient ω. As a key to control matrix splitting, much more efforts have been devoted on the choice of ω to guarantee the iteration convergence. To strengthen the precoding performance, the Gauss-Seidel (GS) iteration which entails a faster convergence than both the Jacobi and Richardson iterations is adopted in [27], [28], and [29]. Furthermore, by incorporating the relaxation coefficient ω to the GS iteration, the successive over-relaxation (SOR) iteration is conceived for improved precoding performance [30], [31]. Furthermore, other works investigating low-complexity iterative precoding algorithms have been explored under different system settings, e.g., [32], [33], [34], [35], [36], and [37].

Apart from these iterative precoding methods, a statistic precoding scheme referred as to randomized iterative precoding algorithm (RIPA) is proposed in [38], which seamlessly incorporate random sampling with iterations. More specifically, RIPA not only ensures a low complexity cost due to its exponentially fast convergence, but also exhibits global convergence in more general settings. This significantly broadens the flexibility of the low-complexity linear precoding,

Manuscript received 11 June 2023; revised 15 December 2023 and 10 May 2024; accepted 10 June 2024. Date of publication 3 July 2024; date of current version 11 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62371124, Grant 61720106003, Grant 62225107, Grant 62201145, Grant 62231019, Grant 61902149, and Grant 62311530098; in part by the Fundamental Research Funds for the Central Universities under Grant 2242022k60002; and in part by the National Key Research and Development Program of China under Grant 2023YFC2205501. The associate editor coordinating the review of this article and approving it for publication was V. Raghavan. *(Corresponding authors: Zheng Wang; Yongming Huang.)*

Zheng Wang, Le Liang, Yili Xia, and Yongming Huang are with the National Mobile Communications Research Laboratory and the Frontiers Science Center for Mobile Information Communication and Security, Southeast University, Nanjing 210096, China (e-mail: wznuaa@gmail.com; huangym@seu.edu.cn).

Shanxiang Lyu is with the College of Cyber Security, Jinan University, Guangzhou 510632, China.

Derrick Wing Kwan Ng is with the School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, NSW 2052, Australia (e-mail: w.k.ng@unsw.edu.au).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2024.3419137.

Digital Object Identifier 10.1109/TWC.2024.3419137

making it promising for applications in various B5G and 6G wireless communication networks. For a better precoding trade-off between the complexity and performance, the modified randomized iterative precoding algorithm (MRIPA) is also presented in [38]. Besides, other randomized iterative precoding schemes can also be found in [39] and [40], but the related convergence analysis about the global convergence is not addressed in those works that call for further investigation.

Despite the aforementioned works that seek a low-complexity implementation of matrix inversion in linear precoding, the complexity reduction of matrix multiplication have not been thoroughly considered. Specifically, the required operations of the matrix multiplication are usually deemed as a preprocessing stage serving for the subsequent low-complexity iterative methods to avoid matrix inversion. Consequently, the computational burden of the matrix multiplication is not explicitly accounted in many existing low-complexity linear precoding schemes. In fact, compared to matrix inversion with a complexity order of $O(K^3)$, matrix multiplication in linear precoding with a complexity order of $O(NK^2)$ is also computationally expensive [41], especially in scenarios with a large number of antennas. As such, reducing the complexity of matrix multiplication in linear precoding is equally important as the one in matrix inversion. The low-complexity work in [19] proposes an efficient method to approximate the matrix multiplication by interpolation but without concerning the matrix inversion. To address this fundamental issue, an efficient randomized iterative precoding algorithm (ERIPA) is proposed in this paper, which aims to efficiently bypass the need for both matrix inversion and matrix multiplication in linear precoding at a low complexity cost. To the best of our knowledge, this is the first comprehensive study on the complexity reduction of matrix multiplication in linear precoding.

In summary, we advance the research of low-complexity iterative precoding in the following several aspects.

- First of all, the efficient randomized iterative precoding algorithm (ERIPA) is designed, which approximates both matrix inversion and matrix multiplication in linear precoding via a statistic approach.
- Secondly, the convergence analysis with respect to ERIPA is carried out and we demonstrate its exponentially fast and global convergence to the target solution. Upon the result, the convergence rate of ERIPA is explicitly derived, and the impacts of the sampling distribution choices on the randomized iterations are also investigated.
- Thirdly, by customizing the conditional sampling into ERIPA, remarkable convergence gain can be achieved, which leads to an enhanced convergence performance while maintaining a lower complexity cost.

The rest of this paper is organized as follows. Section II presents the classic linear precoding for downlink massive MIMO and briefly reviews the existing low-complexity iterative precoding schemes. In Section III, ERIPA is proposed with the complexity analysis for efficient implementation. Then, in Section IV, convergence analysis associated with ERIPA is carried out to demonstrate its global and exponential convergence, followed by the investigation of the sampling distribution in randomized iterations of ERIPA. In Section V, the mechanism of conditional sampling is leveraged by ERIPA for striking a better precoding trade-off between performance and complexity. In Section VI, simulations associated with the proposed ERIPA for downlink systems in massive MIMO are presented, and Section VII gives the conclusion in the end.

Notation: Matrix and its column vectors are respectively indicated by upper and lowercase boldface letters. The transpose, conjugate transpose, inverse, square root and pseudoinverse of matrix $\mathbf{B}$ are represented by $\mathbf{B}^T, \mathbf{B}^H, \mathbf{B}^{-1}$, $\mathbf{B}^{\frac{1}{2}}$ and $\mathbf{B}^{\dagger}$. We employ $\mathbf{b}_i$ for the ith column of matrix $\mathbf{B}$, $b_{i,j}$ for the element in the ith row and jth column of matrix $\mathbf{B}$. Next, let $\langle \mathbf{X}, \mathbf{Y} \rangle_{F(\mathbf{W}^{-1})} \triangleq \text{Tr}(\mathbf{X}^H\mathbf{W}^{-1}\mathbf{Y}\mathbf{W}^{-1})$ stand for the weighted Frobenius inner product, where $\mathbf{W} \in \mathbb{C}^{n\times n}$ is a symmetric positive definite matrix and $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{n\times n}$. In addition, let $\|\mathbf{X}\|^2_{F(\mathbf{W}^{-1})} \triangleq \text{Tr}(\mathbf{X}^H\mathbf{W}^{-1}\mathbf{X}\mathbf{W}^{-1}) = \|\mathbf{W}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}^{-\frac{1}{2}}\|^2_F$, where $\|\cdot\|_F$ denotes Frobenius norm with the identity matrix $\mathbf{I}$ and $\text{Tr}(\cdot)$ stands for the matrix trace. $E[\cdot]$ indicates the expectation of a random variable. Finally, in this paper, the computational complexity is measured by the number of multiplication operations.

II. Background

In this section, the classic downlink linear precoding for massive MIMO is first introduced, followed by a brief review of existing low-complexity iterative precoding schemes.

A. Linear Precoding for Downlink Massive MIMO

Assume that the base station (BS) is equipped with N antennas and there are K single-antenna user terminals (UT) being served simultaneously such that $N \geq K$ holds. During the downlink transmission of this multi-user massive MIMO, let $\mathbf{s} = [s_1, \ldots, s_K]$ represent the source information sent to the K users, and then the received signal $\mathbf{y} \in \mathbb{C}^K$ at the UTs stacking in a vector can be expressed by [13], [14], [15], [16], [17], [18], [19], [30], [31], [32], [33], and [34], etc.

$$\mathbf{y} = \sqrt{\varrho}\mathbf{H}^H\mathbf{G}\mathbf{s} + \mathbf{n}. \tag{1}$$

Here, $\mathbf{H} \in \mathbb{C}^{N\times K}$ is the transmission channel matrix, $\mathbf{G} \in \mathbb{C}^{N\times K}$ denotes the desired precoding matrix, the coefficient $\varrho > 0$ represents the transmit power on average at the BS, $\mathbf{n} \in \mathbb{C}^K$ is the additive white Gaussian noise (AWGN) with each element following $\mathcal{CN}(0, \sigma^2)$. Note that $\mathbf{s}$ and $\mathbf{G}$ should satisfy the requirements of $E\{\|\mathbf{s}\|^2\} = 1$ and $\text{Tr}(\mathbf{G}\mathbf{G}^H) = 1$ due to the power constraint.

Given the system model in (1), in order to mitigate the multiuser interference, the linear ZF precoder for the downlink transmission is given by [42]

$$\mathbf{G}_{\text{zf}} = \beta\mathbf{H}(\mathbf{H}^H\mathbf{H})^{-1}, \tag{2}$$

where $\beta \geq 0$ is a scaling coefficient for normalizing the signal power. Clearly, ZF precoder guides the transmitted signal to the desired user but nulling in the undesired transmission directions to others. Moreover, regularized ZF (RZF) precoding is formulated by [43]

$$\mathbf{G}_{\text{rzf}} = \beta\mathbf{H}(\mathbf{H}^H\mathbf{H} + \xi\mathbf{I})^{-1} \tag{3}$$

based on the criterion of mean squared error (MSE), where $\mathbf{I}$ denotes the identity matrix with a scalar regularization factor $\xi \geq 0$. In principle, RZF precoding naturally degenerates to ZF precoding when $\xi \rightarrow 0$ while it works as the traditional maximal ratio transmission (MRT) if $\xi \rightarrow \infty$. Note that the instantaneous channel state information (CSI) is considered for the linear precoding [13], [14], [15], [16], [17], [18], [19], [30], [31], [32], [33], [34] etc..

B. Low-Complexity Iterative Precoding

In [22], the Neumann series (NS) is employed to facilitate the design of low-complexity linear precoding via circumventing the need for matrix inversion, thus reducing computational complexity. Specifically, by letting the precoding matrix

$$\mathbf{G} = \beta \mathbf{H} \mathbf{A}^{-1} \tag{4}$$

with $\mathbf{A} = \mathbf{H}^H \mathbf{H} \in \mathbb{C}^{K \times K}$ for ZF precoding and $\mathbf{A} = \mathbf{H}^H \mathbf{H} + \xi \mathbf{I} \in \mathbb{C}^{K \times K}$ for RZF precoding, the Neumann series approximates the matrix inversion of $\mathbf{A}$ by the following series expansion (SE) [11]

$$\mathbf{A}^{-1} = \sum_{k=0}^{\infty} (\mathbf{I} - \mathbf{\Theta} \mathbf{A})^k \mathbf{\Theta} \tag{5}$$

if the convergence condition $\lim_{k \rightarrow \infty} (\mathbf{I} - \mathbf{\Theta} \mathbf{A})^k = \mathbf{0}$ is satisfied.[1] Here, $\mathbf{\Theta}$ represents a $K \times K$ diagonal matrix, k is the index of the iterations. Other modifications based on the SE for low-complexity linear precoding are referred to [18], [21], [44], and [45]. Based on the approximated matrix inversion obtained by SE, a dynamic updating mechanism for computing the matrix inversion given any small variations of the system configuration (e.g., add or remove a user terminal from the network) is also given in [46]. Nevertheless, all these SE-based low-complexity iterative precoding methods impose the same stringent convergence requirement as the NS.

On the other hand, given (4), the system model in (1) becomes

$$\mathbf{y} = \sqrt{\varrho} \mathbf{H}^H \mathbf{G} \mathbf{s} + \mathbf{n} = \sqrt{\varrho} \beta \mathbf{H}^H \mathbf{H} \mathbf{A}^{-1} \mathbf{s} + \mathbf{n} = \sqrt{\varrho} \beta \mathbf{H}^H \mathbf{H} \mathbf{t} + \mathbf{n}, \tag{6}$$

where $\mathbf{t} = \mathbf{A}^{-1} \mathbf{s} \in \mathbb{C}^K$. In this approach, the problem of matrix inversion regarding $\mathbf{A}$ is equivalently transformed into the problem of finding $\mathbf{t}$ in the following linear system:

$$\mathbf{A} \mathbf{t} = \mathbf{s}, \tag{7}$$

since they output the same result by $\mathbf{G} \mathbf{s} = \beta \mathbf{H} \mathbf{A}^{-1} \mathbf{s} = \beta \mathbf{H} \mathbf{t}$. Therefore, letting the target precoding solution be $\mathbf{t}^* = \mathbf{A}^{-1} \mathbf{s}$, the classic iterative methods can be introduced to solve the linear system in (7) at a low complexity cost [27].

Specifically, the matrix splitting method with respect to $\mathbf{A} = \mathbf{P} + \mathbf{Q}$, with $\mathbf{P} \in \mathbb{C}^{K \times K}$ and $\mathbf{Q} \in \mathbb{C}^{K \times K}$, is applied. Based on it, iterative methods can be employed to compute the target $\mathbf{t}^*$ via the following iterations [47]

$$\mathbf{t}^{k+1} = \mathbf{B} \mathbf{t}^k + \mathbf{f}, \tag{8}$$

with the *iteration matrix* $\mathbf{B} = -\mathbf{P}^{-1} \mathbf{Q} = \mathbf{I} - \mathbf{P}^{-1} \mathbf{A} \in \mathbb{C}^{K \times K}$ and $\mathbf{f} = \mathbf{P}^{-1} \mathbf{s} \in \mathbb{C}^K$. In theory, the convergence of the iterations is ensured if [24]

$$\lim_{k \rightarrow \infty} \mathbf{B}^k = \mathbf{0}, \tag{9}$$

where k stands for the number of iterations.

From (8), the crux of these iterative methods clearly lies on the way of matrix splitting, which naturally accounts for different convergence performance. Typically, for the Jacobi iteration method, the formation of matrix splitting with $\mathbf{P} = \mathbf{D}$, $\mathbf{Q} = \mathbf{L} + \mathbf{U}$ is employed, where $\mathbf{D} \in \mathbb{C}^{K \times K}$, $\mathbf{L} \in \mathbb{C}^{K \times K}$, $\mathbf{U} \in \mathbb{C}^{K \times K}$ are diagonal, lower triangular, upper triangular entries of matrix $\mathbf{A}$. Besides, in Richardson iteration, matrix splitting with $\mathbf{P} = \frac{1}{\omega} \mathbf{I}$ and $\mathbf{Q} = \mathbf{A} - \frac{1}{\omega} \mathbf{I}$ is applied, and the coefficient $\omega > 0$ serves as a relaxation coefficient. As for the GS method, its iteration is accomplished with that of the matrix splitting $\mathbf{P} = \mathbf{D} + \mathbf{U}$ and $\mathbf{Q} = \mathbf{L}$, which achieves a faster convergence performance than Jacobi and Richardson iterations [29]. Furthermore, based on the GS iteration, the relaxation coefficient $1 < \omega < 2$ is introduced so as to the SOR iteration conceived by

$$(\mathbf{D} + \omega \mathbf{U}) \mathbf{t}^{k+1} = [(1 - \omega) \mathbf{D} - \omega \mathbf{L}] \mathbf{t}^k + \omega \mathbf{s}, \tag{10}$$

where the GS iteration turns out to be a special case of SOR iteration with $\omega = 1$. Apart from these deterministic iterative methods, the statistical iterative methods resorted to random sampling are also introduced into the linear precoding, which are briefly described in the following. Note that the demand of near linear precoding performance with less complexity cost always holds independent of the accuracy of the obtained CSI.

C. Randomized Iterative Precoding

In [38], randomized iterative precoding algorithm (RIPA) is presented for downlink systems in massive MIMO. By merging the random sampling to the traditional iterations, RIPA attains a low computational complexity for precoding while enjoying the globally exponential convergence at the same time.

In particular, RIPA operates via the following iterations

$$\mathbf{t}^{k+1} = \mathbf{t}^k + \mathbf{A}^{-1} \mathbf{S}_k (\mathbf{S}_k^H \mathbf{S}_k)^{-1} \mathbf{S}_k^H (\mathbf{s} - \mathbf{A} \mathbf{t}^k). \tag{11}$$

At each iteration, the matrix $\mathbf{S}_k \in \mathbb{C}^{K \times q}$ is sampled randomly over an artificially designed sampling distribution $\mathcal{D}(\mathbf{S}_k)$, namely, $\mathbf{S}_k \sim \mathcal{D}(\mathbf{S}_k)$. More precisely, in RIPA, the sampling set of $\mathbf{S}_k$ is designed as $\{\mathbf{I}_{:,\mathcal{Q}_1}, \ldots, \mathbf{I}_{:,\mathcal{Q}_r}\}$, where $\mathbf{I}_{:,\mathcal{Q}_i}$, $1 \leq i \leq r$ indicates a column segmentation of $\mathbf{I}_{K \times K}$. Note that $\mathcal{Q}_i \triangleq \{\text{index } 1, \ldots, \text{index } q_i\} \subseteq \{1, \ldots, K\}$ represents a set of index with set size $|\mathcal{Q}_i| = q_i$, i.e.,

$$\mathbf{I}_{:,\mathcal{Q}_i} = [\mathbf{I}_{:,\text{index } 1}, \ldots, \mathbf{I}_{:,\text{index } q_i}], \tag{12}$$

which forms a block like the following example:

$$\underbrace{\{1, 2, 5\}}_{\mathcal{Q}_1} \cup \ldots \cup \underbrace{\{4, 8, 12\}}_{\mathcal{Q}_r} = \{1, \ldots, K\} \tag{13}$$

with $\mathcal{Q}_i \cap \mathcal{Q}_j = \emptyset$, $1 \leq i \neq j \leq r$. More specifically, the allocation in every set $\mathcal{Q}_i$ are decided at the beginning such that the following sampling regarding $\mathbf{S}_k$ is performed

[1] In [21], the sufficient convergence condition is derived explicitly as $N/K \geq 5.83$.

iteratively. In addition, the same block size $|\mathcal{Q}_1| = \ldots = |\mathcal{Q}_r| = K/r = q$, namely, $q_1 = \ldots = q_r = q$, is employed for simplicity. In the following, we reveal the convergence of RIPA.

Theorem 1 ([38]): Regarding the downlink massive MIMO, let $\mathbf{S}_k \in \{\mathbf{I}_{:,\mathcal{Q}_1}, \ldots, \mathbf{I}_{:,\mathcal{Q}_r}\}$ *be obtained by sampling from distribution* $\mathcal{D}(\mathbf{S}_k)$*, RIPA following (11) converges by*

$$E[\|\mathbf{t}^k - \mathbf{t}^*\|^2_{F(\mathbf{A}^H\mathbf{A})}] \leq \rho^k \|\mathbf{t}^0 - \mathbf{t}^*\|^2_{F(\mathbf{A}^H\mathbf{A})} \tag{14}$$

with global and exponential convergence rate

$$\rho < 1. \tag{15}$$

From Theorem 1, it is easy to confirm the global convergence because the associated convergence rate ρ of the iteration in RIPA is always less than 1 regardless of any specific requirements. Then, based on RIPA, by beneficially taking advantages of the conditional sampling probability, MRIPA is designed for further improvements about iteration efficiency and convergence [38].

III. Efficient Randomized Iterative Precoding Algorithm

As clearly shown in (2) and (3), the computations of linear precoding consist of both matrix multiplication and matrix inversion operations, e.g.,

$$\mathbf{G}_{\text{zf}} = \beta\mathbf{H}\overbrace{(\underbrace{\mathbf{H}^H\mathbf{H}}_{\text{matrix multiplication}})^{-1}}^{\text{matrix inversion}}, \tag{16}$$

where the associated computation is generally expensive with the increment of the system dimension. However, most related low-complexity linear precoding works only focus on bypassing the matrix inversion step while overlooking the necessary complexity reduction regarding to matrix multiplication.

Specifically, as shown in (4) and (7), the existing complexity reductions for matrix inversion are carried out based on the given matrix $\mathbf{A} = \mathbf{H}^H\mathbf{H}$ without concerning the acquisition of $\mathbf{A}$ via matrix multiplication. In other words, they deem the matrix multiplication of $\mathbf{A} = \mathbf{H}^H\mathbf{H}$ as a preprocessing operation, which does not exactly count towards the complexity evaluation. However, the complexity of matrix multiplication $\mathbf{H}^H\mathbf{H}$ (i.e., $O(NK^2)$) is actually comparable to that of matrix inversion $(\mathbf{H}^H\mathbf{H})^{-1}$ (i.e., $O(K^3)$), which is also worthy of being properly reduced. Hence, ERIPA is proposed in this paper, which returns the desired solution of linear precoding by efficiently approximating both the matrix inversion and multiplication. By doing so, significant computational complexity can be saved without compromising the rapid convergence of randomized iterative precoding.

A. Algorithm Design

To start with, we focus on the low-complexity ZF precoding for downlink systems in massive MIMO in (1). Then, the randomized iteration in ERIPA is designed by

$$\mathbf{t}^{k+1} = \mathbf{t}^k + \mathbf{I}_{:,\mathcal{Q}_i}(\mathbf{H}^H_{:,\mathcal{Q}_i}\mathbf{H}_{:,\mathcal{Q}_i})^{-1}(\mathbf{I}^H_{:,\mathcal{Q}_i}\mathbf{s} - \mathbf{H}^H_{:,\mathcal{Q}_i}\mathbf{H}\mathbf{t}^k), \tag{17}$$

while at each iteration the set of column indexes $\mathcal{Q}_i$ is sampled from the distribution $\mathcal{D}$ with a sampling probability

$$p_i \triangleq \mathcal{D}(\mathcal{Q}_i), \tag{18}$$

such that the matrix $\mathbf{H}_{:,\mathcal{Q}_i} = \mathbf{H}\mathbf{I}_{:,\mathcal{Q}_i} \in \mathbb{C}^{N\times q}$ can be attained with the sampling result $\mathcal{Q}_i$. [2] For example, if the index set $\mathcal{Q}_1 = \{1, 2, 3\}$ is selected, then we have $\mathbf{H}_{:,\mathcal{Q}_1} = [\mathbf{h}_1\ \mathbf{h}_2\ \mathbf{h}_3]$, where $\mathbf{h}_i$ denotes the i-th column of channel matrix $\mathbf{H}$.

Algorithm 1 ERIPA for Downlink Precoding in Massive MIMO

```
Require: $\mathbf{H}$, $\mathbf{t}^0 = \mathbf{0}$, $L$, $\beta$, $\xi$
Ensure: Near ZF or RZF precoding result $\mathbf{G}\mathbf{s} = \beta\mathbf{H}\mathbf{t}^L$
for $k = 1, \ldots, L$ do
    for $m = 1, \ldots, K/q$ do
        Sample $\mathcal{Q}_i$ by (18)
        Iterate $\mathbf{t}$ according to (17) or (19)
    end for
end for
Output $\mathbf{G}\mathbf{s} = \beta\mathbf{H}\mathbf{A}^{-1}\mathbf{s} = \beta\mathbf{H}\mathbf{t}^L$
```

Similarly, as for the near RZF precoding performance, the randomized iteration in the proposed ERIPA becomes[3]

$$\begin{aligned}\mathbf{t}^{k+1} = \mathbf{t}^k + \mathbf{I}_{:,\mathcal{Q}_i}(\mathbf{H}^H_{:,\mathcal{Q}_i}\mathbf{H}_{:,\mathcal{Q}_i} + \mathbf{I}^H_{:,\mathcal{Q}_i}\xi\mathbf{I}_{:,\mathcal{Q}_i})^{-1}[\mathbf{I}^H_{:,\mathcal{Q}_i}\mathbf{s}\\ - (\mathbf{H}^H_{:,\mathcal{Q}_i}\mathbf{H} + \mathbf{I}^H_{:,\mathcal{Q}_i}\xi\mathbf{I})\mathbf{t}^k].\end{aligned} \tag{19}$$

Here, the same block size $|\mathcal{Q}_1| = \ldots = |\mathcal{Q}_r| = K/r = q$ is exploited while the indices contained in $\mathcal{Q}_i$, $1 \leq i \leq r$ are fixed at the beginning. Intuitively, a sequential order such as $\mathbf{H}_{:,\mathcal{Q}_i} = [\mathbf{h}_{(i-1)q+1}, \ldots, \mathbf{h}_{iq}]$ with $\mathcal{Q}_i = \{(i-1)q+1, \ldots, iq\}$, can be employed to establish these index sets. Please note that the random sampling based on (18) is carried out among the sets $\{\mathcal{Q}_1, \ldots, \mathcal{Q}_r\}$ rather than the indexes within a specific set $\mathcal{Q}_i$. In addition, we set $\mathbf{t}^0 = \mathbf{0}$ for the sake of simplicity. Consequently, according to the random selection of $\mathcal{Q}_i$ at each iteration, the random iteration in (17) or (19) is carried out block-by-block given the block size q, which converges to the desired ZF precoding solution $\mathbf{t}^* = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{s}$ or RZF precoding solution $\mathbf{t}^* = (\mathbf{H}^H\mathbf{H} + \xi\mathbf{I})^{-1}\mathbf{s}$ exponentially fast.

Clearly, given the random iterations in (17) and (19), the complicated matrix computations, such as the multiplication of $\mathbf{H}^H\mathbf{H}$ and the inversion of $(\mathbf{H}^H\mathbf{H})^{-1}$ are not involved in the proposed ERIPA. Here, we point out that a small choice $q \ll N$ is recommended to simplify the matrix operation involving $\mathbf{H}_{:,\mathcal{Q}_i}$ in (17). However, a large value of q also facilitates a better convergence performance as more elements of $\mathbf{t}$ are updated at a time. Therefore, there is a trade-off with respect to q between the convergence and the computational complexity in the proposed ERIPA. As for the current work, the flexible choice of $1 < q \leq \sqrt{K}$ is preferred, and further study regarding to q is one of our future work.

[2] More specifically, the set of column indexes $\mathcal{Q}_i$ behaves like a random variable with distribution $\mathcal{D}(\mathcal{Q}_i)$, so that $\mathcal{Q}_i \sim \mathcal{D}(\mathcal{Q}_i)$, where $\sim$ means "with a probability distribution given by".

[3] Clearly, the randomized iteration for RZF precoding degenerates to ZF precoding when ξ is 0.

B. Efficient Implementation by Complexity Analysis

Since the computations about $\mathbf{I}_{:,\mathcal{Q}_i}^H \xi \mathbf{I}_{:,\mathcal{Q}_i}$ and $\mathbf{I}_{:,\mathcal{Q}_i}^H \xi \mathbf{t}^k$ in (19) can be obtained instantly, we consider the complexity of ERIPA with respect to the random iteration in (17), which could be significantly simplified by taking advantages of the special structure of $\mathbf{I}_{:,\mathcal{Q}_i}$.

Typically, given the index set $\mathcal{Q}_i$ with a sequential order, one can express the matrix $\mathbf{I}_{:,\mathcal{Q}_i}^H$ by

$$\mathbf{I}_{:,\mathcal{Q}_i}^H = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & 0 & \ddots & \ddots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots & \vdots & \ddots & \ddots & 0 & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}, \tag{20}$$

which means both $\mathbf{I}_{:,\mathcal{Q}_i}^H$ and $\mathbf{I}_{:,\mathcal{Q}_i}$ are controlled by the inner $q \times q$ nonzero submatrix. More importantly, each column or row of such a submatrix only contains one nonzero element 1, thus leading to a further complexity reduction in the proposed ERIPA. More specifically, thanks to the special construction of matrix $\mathbf{I}_{:,\mathcal{Q}_i}$ in (20), multiplying $\mathbf{I}_{:,\mathcal{Q}_i}$ with $(\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i})^{-1}$ in (17) could be simplified by the direct shift as well as matrix augmentation by 0, e.g.,

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ g & h & i \\ 0 & 0 & 0 \\ d & e & f \\ a & b & c \end{bmatrix}, \tag{21}$$

which results in a negligible complexity. Meanwhile, similar efficient operations can also be employed to obtain the terms $\mathbf{I}_{:,\mathcal{Q}_i}^H \mathbf{s}$ and $\mathbf{H}_{:,\mathcal{Q}_i}$ in (17).

Since the $K \times q$ matrix $\mathbf{I}_{:,\mathcal{Q}_i}(\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i})^{-1}$ is completely controlled by the $q \times q$ nonzero submatrix in it (i.e., all the other elements excluding this submatrix are 0s), the complexity of multiplying it with the vector term $(\mathbf{I}_{:,\mathcal{Q}_i}^H \mathbf{s} - \mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}\mathbf{t}^k) \in \mathbb{C}^q$ can be simplified as q^2. As for the term $\mathbf{H}\mathbf{t}^k$, it can be obtained with computational complexity NK. However, since only q elements in $\mathbf{t}$ are updated at each iteration, given $\mathbf{t}^0 = \mathbf{0}$, the complexity of computing $\mathbf{H}\mathbf{t}^k$ at each iteration can be reduced to qN. Furthermore, multiplying $\mathbf{H}_{:,\mathcal{Q}_i}^H$ with $\mathbf{H}\mathbf{t}^k$ costs qN. To sum up, considering the fact that the computational complexity of $(\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i})^{-1}$ is $q^3 + q^2 N$, the total complexity for each iteration in the proposed ERIPA is $q^3 + q^2 N + q^2 + 2qN$, which is $O(KN)$ with the block size $q = \sqrt{K}$.

On the other hand, to ensure a fair comparison over other iterative precoding algorithms (i.e., all the components of $\mathbf{t}$ are updated at each instant), a *full iteration* that involves K/q iterations of (17) is also used by ERIPA.[4] Hence, this results in the complexity $O(K^{1.5}N)$ of a full iteration in ERIPA, which is still superior to other low-complexity iterative precoding algorithms. To sum up, the proposed ERIPA for the downlink systems in massive MIMO is shown in Algorithm 1. In particular, the outer loop given by steps 1-6 of Algorithm 1 represents a full iteration of ERIPA and it includes K/q times randomized iterations of (17) or (19). Meanwhile, for a better understanding, in Table I we contrast the computational complexity of the proposed ERIPA to those of the related literature, in which k represents the number of iterations fully implemented.

IV. Convergence Studies

In this section, we analyze the convergence behaviour of the proposed ERIPA. By means of the MSE criterion, the global and exponentially fast convergence of ERIPA is proved based on the derived convergence rate. Meanwhile, the selection of the sampling distribution $\mathcal{D}$ is also investigated to harness the randomized iteration in ERIPA.

A. Global and Exponential Convergence

First of all, given the full column rank property of the channel matrix $\mathbf{H}$, the matrix $\mathbf{H}^H\mathbf{H}$ turns out to be symmetric positive definite, which can be easily verified in the sense of lattice theory [48]. From it, the eigenvector matrix $\mathbf{V}$ of $\mathbf{H}^H\mathbf{H}$ can be readily obtained as (e.g., [49], Chapter 8)

$$\mathbf{V}\mathbf{V} = \mathbf{V}\mathbf{V}^H = \mathbf{V}^H\mathbf{V} = \mathbf{H}^H\mathbf{H}. \tag{22}$$

We emphasize that $\mathbf{V}$ is not involved in the implementation of ERIPA, but is only exploited to pave the way for the convergence analysis.

Theorem 2: Regarding the downlink massive MIMO, let $\mathbf{S}_k = \mathbf{H}_{:,\mathcal{Q}_i}$ be obtained by sampling from the distribution $\mathcal{D}$, ERIPA given in (17) is convergent

$$E[\|\mathbf{V}(\mathbf{t}^k - \mathbf{t}^*)\|^2] \leq \rho^k \|\mathbf{V}(\mathbf{t}^0 - \mathbf{t}^*)\|^2 \tag{23}$$

with an exponentially fast and global convergence rate

$$\rho = 1 - \lambda_{\text{min}}(\mathbf{H}^H\mathbf{H}E[\mathbf{Z}]) < 1, \tag{24}$$

where

$$\mathbf{Z} \triangleq \mathbf{I}_{:,\mathcal{Q}_i}(\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i})^{-1}\mathbf{I}_{:,\mathcal{Q}_i}^H. \tag{25}$$

Proof: First of all, based on the definition in (25), the random iteration in (17) becomes

$$\mathbf{t}^{k+1} = \mathbf{t}^k + \mathbf{Z}\mathbf{H}^H\mathbf{H}(\mathbf{t}^* - \mathbf{t}^k), \tag{26}$$

which can be further reformulated as

$$\mathbf{t}^{k+1} - \mathbf{t}^* = (\mathbf{I} - \mathbf{Z}\mathbf{H}^H\mathbf{H})(\mathbf{t}^k - \mathbf{t}^*). \tag{27}$$

Then, for ease of presentation, by letting $\mathbf{r}_k = \mathbf{t}^k - \mathbf{t}^*$, we have

$$\mathbf{r}_{k+1} = (\mathbf{I} - \mathbf{Z}\mathbf{H}^H\mathbf{H})\mathbf{r}_k, \tag{28}$$

and the following derivations can be obtained

$$\begin{aligned} &\|\mathbf{V}(\mathbf{t}^k - \mathbf{t}^*)\|^2 \\ &\quad = \mathbf{r}_k^H \mathbf{V}^H \mathbf{V} \mathbf{r}_k \qquad (29) \\ &\quad \overset{(a)}{=} \mathbf{r}_{k-1}^H (\mathbf{I} - \mathbf{Z}\mathbf{H}^H\mathbf{H})^H \mathbf{V}^H \mathbf{V} (\mathbf{I} - \mathbf{Z}\mathbf{H}^H\mathbf{H}) \mathbf{r}_{k-1} \\ &\quad = \mathbf{r}_{k-1}^H (\mathbf{V} - \mathbf{V}\mathbf{Z}\mathbf{H}^H\mathbf{H})^H (\mathbf{V} - \mathbf{V}\mathbf{Z}\mathbf{H}^H\mathbf{H}) \mathbf{r}_{k-1} \\ &\quad = \mathbf{r}_{k-1}^H (\mathbf{V}^H\mathbf{V} - \mathbf{V}^H\mathbf{V}\mathbf{Z}\mathbf{H}^H\mathbf{H} - \mathbf{H}^H\mathbf{H}\mathbf{Z}^H\mathbf{V}^H\mathbf{V} \\ &\qquad + \mathbf{H}^H\mathbf{H}\mathbf{Z}^H\mathbf{V}^H\mathbf{V}\mathbf{Z}\mathbf{H}^H\mathbf{H}) \mathbf{r}_{k-1} \end{aligned}$$

[4] Note that the values of K/q and q should be integers in practical cases.

TABLE I
COMPLEXITY COMPARISON OF ITERATIVE PRECODING SCHEMES

	Matrix Multiplication	Matrix Inversion		Matrix Multiplication	Matrix Inversion
ZF or RZF	$O(NK^2)$	$O(K^3)$	GS [28]	$O(NK^2)$	$O(K^2 \cdot k)$
NS [22]	$O(NK^2)$	$O(K^2 \cdot k)$ for $k \leq 2$	SOR [31]	$O(NK^2)$	$O(K^2 \cdot k)$
Newton [25]	$O(NK^2)$	$O(K^2 \cdot k)$	RIPA [38]	$O(NK^2)$	$O(K^{2.5} \cdot k)$
Jacobi [20]	$O(NK^2)$	$O(K^2 \cdot k)$	MRIPA [38]	$O(NK^2)$	$O(K^2 \cdot k)$
Richardson [26]	$O(NK^2)$	$O(K^2 \cdot k)$	ERIPA	$O(K^{1.5}N \cdot k)$ in total	

$$\overset{(b)}{=} \mathbf{r}_{k-1}^H(\mathbf{V}^H\mathbf{V} - \mathbf{V}^H\mathbf{V}\mathbf{Z}\mathbf{H}^H\mathbf{H})\mathbf{r}_{k-1}, \tag{30}$$

where the equality in (a) stems from (28) and the equality in (b) holds due to the facts $\mathbf{Z}^H\mathbf{V}^H\mathbf{V}\mathbf{Z} = \mathbf{Z}$ and $\mathbf{V}^H\mathbf{V} = \mathbf{H}^H\mathbf{H}$.

Next, according to *law of total expectation*,[5] we can get

$$E[\|\mathbf{V}\mathbf{r}_k\|^2] = E[E[\|\mathbf{V}\mathbf{r}_k\|^2|\mathbf{r}_{k-1}]]. \tag{31}$$

Meanwhile, given (30), it follows that

$$\begin{aligned}
E[\|\mathbf{V}\mathbf{r}_k\|^2|\mathbf{r}_{k-1}] &= E[\mathbf{r}_{k-1}^H(\mathbf{V}^H\mathbf{V} - \mathbf{V}^H\mathbf{V}\mathbf{Z}\mathbf{H}^H\mathbf{H})\mathbf{r}_{k-1}] \\
&= E[\mathbf{r}_{k-1}^H(\mathbf{V}^H\mathbf{V} - \mathbf{V}^H\mathbf{V}\mathbf{Z}\mathbf{V}^H\mathbf{V})\mathbf{r}_{k-1}] \\
&= E[\mathbf{r}_{k-1}^H\mathbf{V}^H(\mathbf{I} - \mathbf{V}\mathbf{Z}\mathbf{V}^H)\mathbf{V}\mathbf{r}_{k-1}] \\
&= E[\langle(\mathbf{I} - \mathbf{V}\mathbf{Z}\mathbf{V}^H)\mathbf{V}\mathbf{r}_{k-1}, \mathbf{V}\mathbf{r}_{k-1}\rangle] \\
&= \langle(\mathbf{I} - \mathbf{V}E[\mathbf{Z}]\mathbf{V}^H)\mathbf{V}\mathbf{r}_{k-1}, \mathbf{V}\mathbf{r}_{k-1}\rangle \\
&\leq \|\mathbf{I} - \mathbf{V}E[\mathbf{Z}]\mathbf{V}^H\| \cdot \|\mathbf{V}\mathbf{r}_{k-1}\|^2 \\
&\overset{(c)}{=} \lambda_{\text{max}}(\mathbf{I} - \mathbf{V}E[\mathbf{Z}]\mathbf{V}^H)\|\mathbf{V}\mathbf{r}_{k-1}\|^2 \\
&= (1 - \lambda_{\text{min}}(\mathbf{V}E[\mathbf{Z}]\mathbf{V}^H))\|\mathbf{V}\mathbf{r}_{k-1}\|^2 \\
&= (1 - \lambda_{\text{min}}(\mathbf{V}\mathbf{V}^H E[\mathbf{Z}]))\|\mathbf{V}\mathbf{r}_{k-1}\|^2 \\
&= (1 - \lambda_{\text{min}}(\mathbf{H}^H\mathbf{H}E[\mathbf{Z}]))\|\mathbf{V}\mathbf{r}_{k-1}\|^2 \\
&= \rho\|\mathbf{V}\mathbf{r}_{k-1}\|^2
\end{aligned} \tag{32}$$

with

$$\rho = 1 - \lambda_{\text{min}}(\mathbf{H}^H\mathbf{H}E[\mathbf{Z}]), \tag{33}$$

where $\lambda_{\text{min}}(\cdot)$ and $\lambda_{\text{max}}(\cdot)$ respectively represent the minimum and maximum eigenvalues of a matrix. Here, the upper bound shown in (c) originates from the symmetry of the matrix $\mathbf{I} - \mathbf{V}E[\mathbf{Z}]\mathbf{V}^H$. Therefore, by cooperating (32) into (31), we can arrive at

$$\begin{aligned}
E[\|\mathbf{V}(\mathbf{t}^k - \mathbf{t}^*)\|^2] &= E[E[\|\mathbf{V}(\mathbf{t}^k - \mathbf{t}^*)\|^2|\mathbf{r}_{k-1}]] \\
&\leq \rho E[\|\mathbf{V}(\mathbf{t}^{k-1} - \mathbf{t}^*)\|^2] \\
&\leq \cdots \\
&\leq \rho^k E[\|\mathbf{V}(\mathbf{t}^0 - \mathbf{t}^*)\|^2] \\
&= \rho^k\|\mathbf{V}(\mathbf{t}^0 - \mathbf{t}^*)\|^2.
\end{aligned} \tag{34}$$

On the other hand, the symmetric matrix $\mathbf{Z}$ in the expectation formation (i.e., $E[\mathbf{Z}]$) is positive definite because of the full column rank matrix $\mathbf{H}$, namely,

$$\begin{aligned}
&E[\mathbf{Z}] \\
&= \sum_{i=1}^{r} p_i \mathbf{I}_{:,\mathcal{Q}_i}(\mathbf{H}_{:,\mathcal{Q}_i}^H\mathbf{H}_{:,\mathcal{Q}_i})^{-1}\mathbf{I}_{:,\mathcal{Q}_i}^H \\
&= \left(\sum_{i=1}^{r} p_i^{\frac{1}{2}} \mathbf{I}_{:,\mathcal{Q}_i}(\mathbf{H}_{:,\mathcal{Q}_i}^H\mathbf{H}_{:,\mathcal{Q}_i})^{-\frac{1}{2}}(\mathbf{H}_{:,\mathcal{Q}_i}^H\mathbf{H}_{:,\mathcal{Q}_i})^{-\frac{1}{2}}\mathbf{I}_{:,\mathcal{Q}_i}^H p_i^{\frac{1}{2}}\right) \\
&= (\mathbf{I}\mathbf{J})(\mathbf{J}\mathbf{I}^H) \\
&= \mathbf{J}^2
\end{aligned} \tag{35}$$

with the invertible and block diagonal matrix $\mathbf{J} \in \mathbb{C}^{K \times K}$

$$\mathbf{J} = \text{diag}(p_1^{\frac{1}{2}}(\mathbf{H}_{:,\mathcal{Q}_1}^H\mathbf{H}_{:,\mathcal{Q}_1})^{-\frac{1}{2}}, \ldots, p_r^{\frac{1}{2}}(\mathbf{H}_{:,\mathcal{Q}_r}^H\mathbf{H}_{:,\mathcal{Q}_r})^{-\frac{1}{2}}), \tag{36}$$

which leads to the result

$$\lambda_{\text{min}}(E[\mathbf{Z}]) > 0. \tag{37}$$

Meanwhile, since the matrix $\mathbf{H}^H\mathbf{H}$ is symmetric and positive definite as well, we can see that $\lambda_{\text{min}}(\mathbf{H}^H\mathbf{H}E[\mathbf{Z}]) > 0$ yields

$$\rho = 1 - \lambda_{\text{min}}(\mathbf{H}^H\mathbf{H}E[\mathbf{Z}]) < 1, \tag{38}$$

which completes the proof. ■

By Theorem 2, the proposed ERIPA with randomized iteration in (17) exhibits exponentially fast convergence towards the target solution $\mathbf{t}^*$. More importantly, it is clear to see that ERIPA is always convergent as long as $\rho < 1$, making it highly flexible to different practical scenarios of massive MIMO systems including the case of multi-user with multiple antennas.[6] Similarly, the convergence proof for the randomized iteration in (19) converging to RZF precoding can be readily attained in the same way, which is omitted here for brevity. On the other hand, to ensure the approximation error smaller than a specific value

$$E[\|\mathbf{V}(\mathbf{t}^k - \mathbf{t}^*)\|^2] \leq \epsilon\|\mathbf{V}(\mathbf{t}^0 - \mathbf{t}^*)\|^2 \tag{39}$$

with $0 < \epsilon < 1$, the iteration number k is lower bounded by

$$k \geq \frac{1}{1-\rho}\log\left(\frac{1}{\epsilon}\right), \tag{40}$$

[5] The expected value of the conditional expected value of random variable X given random variable Y is the same as the expected value of X, i.e., $E(X) = E(E(X|Y))$.

[6] Here, even the imperfect channel matrix $\widehat{\mathbf{H}}$ is encountered, the convergence of ERIPA still holds as the matrix $\widehat{\mathbf{H}}^H\widehat{\mathbf{H}}$ remains symmetric positive definite.

with inequality $\ln(1-\delta) < -\delta$ for $0 < \delta < 1$. Intuitively, a tractable ERIPA can be attained by tuning k. In addition, as shown in (23), the convergence of randomized iteration can also be beneficially exploited with a better choice of $\mathbf{t}^0$. Therefore, if it is possible, a closer $\mathbf{t}^0$ to the target $\mathbf{t}^*$ is highly desired to the proposed ERIPA.

B. Impact of Sampling Mechanism

On the other hand, regarding to the sampling probability $\mathcal{D}$ in (18), a straightforward choice is to exploit the uniform distribution and its impact to the convergence of ERIPA is summarized in the following Corollary.

Corollary 1: Let $\mathcal{Q}_i$ follow the uniform sampling probability

$$p_i = \frac{1}{r}, \tag{41}$$

ERIPA is convergent by

$$E[\|\mathbf{V}(\mathbf{t}^k - \mathbf{t}^*)\|^2] \leq \rho_{\text{uniform}}^k \|\mathbf{V}(\mathbf{t}^0 - \mathbf{t}^*)\|^2 \tag{42}$$

with

$$\rho_{\text{uniform}} = 1 - \frac{1}{r} \cdot \min_i \left\{ \frac{\lambda_{\min}(\mathbf{H}^H\mathbf{H})}{\lambda_{\max}(\mathbf{H}^H_{:,\mathcal{Q}_i}\mathbf{H}_{:,\mathcal{Q}_i})} \right\} < 1. \tag{43}$$

Proof: First of all, based on (35), it follows

$$\lambda_{\min}(\mathbf{H}^H\mathbf{H}E[\mathbf{Z}]) = \lambda_{\min}(\mathbf{H}^H\mathbf{H}\mathbf{J}^2) \overset{(d)}{\geq} \lambda_{\min}(\mathbf{H}^H\mathbf{H})\lambda_{\min}(\mathbf{J}^2), \tag{44}$$

where the inequality in (d) stems from $\lambda_{\min}(\mathbf{E}\mathbf{F}) \geq \lambda_{\min}(\mathbf{E})\lambda_{\min}(\mathbf{F})$ for any positive definite matrices $\mathbf{E}$, $\mathbf{F}$. Then, from (36), we have

$$\lambda_{\min}(\mathbf{J}^2) = \frac{1}{r} \cdot \frac{1}{\lambda_{\max}(\mathbf{H}^H_{:,\mathcal{Q}_i}\mathbf{H}_{:,\mathcal{Q}_i})} \tag{45}$$

with a uniform sampling distribution $p_i = 1/r$.

Next, by (44) and (45), the convergence rate ρ in (24) is derived as

$$\begin{aligned}
\rho &= 1 - \lambda_{\min}(\mathbf{H}^H\mathbf{H}E[\mathbf{Z}]) \\
&\leq 1 - \lambda_{\min}(\mathbf{H}^H\mathbf{H})\lambda_{\min}(\mathbf{J}^2) \\
&\leq 1 - \frac{1}{r} \cdot \min_i \left\{ \frac{\lambda_{\min}(\mathbf{H}^H\mathbf{H})}{\lambda_{\max}(\mathbf{H}^H_{:,\mathcal{Q}_i}\mathbf{H}_{:,\mathcal{Q}_i})} \right\} \\
&= \rho_{\text{uniform}} \\
&< 1,
\end{aligned} \tag{46}$$

where the global convergence $\rho_{\text{uniform}} < 1$ is still achieved as all the eigenvalues of a symmetric and positive definite matrix (i.e., $\mathbf{H}^H\mathbf{H}$) are positive. ■

From (43), on one hand, the convergence rate partially relies on the specific size of q (i.e., $q = K/r = |\mathcal{Q}_i|$) and the specific partitions of $\mathbf{H}$ (i.e., $\mathbf{H}_{:,\mathcal{Q}_i}$). On the other hand, the convergence performance of ρ_{uniform} is also controlled by the condition number $\kappa = \lambda_{\max}(\mathbf{H}^H\mathbf{H})/\lambda_{\min}(\mathbf{H}^H\mathbf{H})$, and a small κ beneficially enables a faster convergence rate.

To be more specific, Fig. 1 is given to illustrate the convergence rate ρ_{uniform} of ERIPA in different antenna sizes of massive MIMO systems, where the sequential partition like $\mathbf{H}_{:,\mathcal{Q}_i} = [\mathbf{h}_{(i-1)q+1}, \ldots, \mathbf{h}_{iq}]$ with $\mathcal{Q}_i = \{(i-1)q+1, \ldots, iq\}, 1 \leq i \leq r$ is applied. In particular, with the increase of $q = 2, 4, 8, 16, 32$, all the convergence rates ρ_{uniform} become faster. This can be interpreted as more elements of $\mathbf{t}$ are updated at a time, thereby leading to a better convergence. Meanwhile, with the increase of antennas at the BS, i.e., $N = 128, 256, 512, 1024$, the convergence rate ρ_{uniform} also becomes faster gradually. This is because the channel matrix $\mathbf{H} \in \mathbb{C}^{N\times K}$ becomes more and more orthogonal such that the off-diagonal elements of $\mathbf{H}^H\mathbf{H}$ become negligible gradually because $\mathbf{h}_i^H * \mathbf{h}_j \rightarrow 0$ (this is known as the effect of *favorable propagation* [50], [51]), which corresponds to a smaller condition number κ of gram matrix $\mathbf{H}^H\mathbf{H}$.

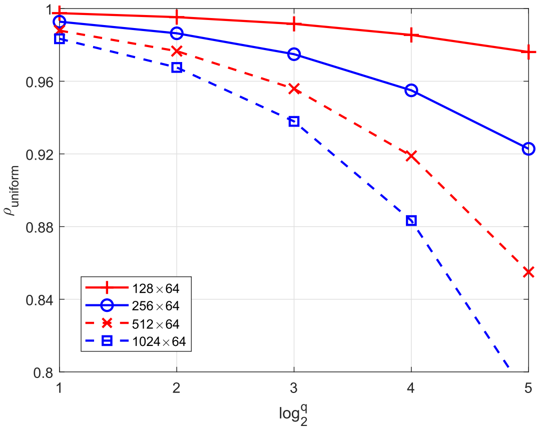


Fig. 1. Convergence comparison of ρ_{uniform} in ERIPA for 128×64, 256×64, 512×64 and 1024×64 massive MIMO systems.

V. Further Enhancement

In this section, by well exploiting the conditional sampling, we show that the proposed randomized iteration in ERIPA turns out to be deterministic, which results in further improvements in both iteration convergence and implementation efficiency.

A. Adoption by Conditional Sampling

Specifically, given the random sampling probability p_i in (18), we define its conditional counterpart as

$$\overline{p}_i \triangleq \mathcal{D}(\mathcal{Q}_i|\mathcal{Q}_j), \ i \neq j = \frac{p_i}{1-p_j}, \ i \neq j. \tag{47}$$

Here, $\mathcal{Q}_i$ and $\mathcal{Q}_j$ respectively denote the sampling results in the current and last randomized iterations. By removing $\mathcal{Q}_j$ (i.e., for iteration $k-1$) from the sampling space of $\mathcal{Q}_i$ (i.e., for iteration k), the case that a same set $\mathcal{Q}_i$ is chosen by two consecutive iterations can be avoided, leading to an improved sampling diversity. As such, the components update of $\mathbf{t}$ can be realized more efficiently, so as to achieve a better convergence.

Specifically, given the conditional sampling probability $\overline{p}_i$ in (47), the conditional expectation about $\mathbf{Z}$ given the previous sampling option $\mathcal{Q}_j$ becomes

$$\begin{aligned}
&E[\mathbf{Z}|\mathcal{Q}_j] \\
&= \sum_{i=1, i\neq j}^{r} \overline{p}_i \mathbf{I}_{:,\mathcal{Q}_i}(\mathbf{H}^H_{:,\mathcal{Q}_i}\mathbf{H}_{:,\mathcal{Q}_i})^{-1}\mathbf{I}^H_{:,\mathcal{Q}_i}
\end{aligned}$$

$$= \left(\sum_{i=1, i \neq j}^{r} \overline{p}_i^{\frac{1}{2}} \mathbf{I}_{:,\mathcal{Q}_i} (\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i})^{-\frac{1}{2}} (\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i})^{-\frac{1}{2}} \mathbf{I}_{:,\mathcal{Q}_i}^H \overline{p}_i^{\frac{1}{2}} \right)$$

$$= (\overline{\mathbf{I}}\overline{\mathbf{J}})(\overline{\mathbf{J}}\overline{\mathbf{I}}^H), \tag{48}$$

and is symmetric and positive definite. More specifically, matrix $\overline{\mathbf{I}} \in \mathbb{C}^{K \times (K-q)}$ accounts for identity matrix $\mathbf{I}$ getting rid of the columns belonging to $\mathcal{Q}_j$, and matrix $\overline{\mathbf{J}} \in \mathbb{C}^{(K-q) \times (K-q)}$ has the formation $\overline{\mathbf{J}} = \text{diag}(\overline{p}_1^{\frac{1}{2}} (\mathbf{H}_{:,\mathcal{Q}_1}^H \mathbf{H}_{:,\mathcal{Q}_1})^{-\frac{1}{2}}, \ldots, \overline{p}_{j-1}^{\frac{1}{2}} (\mathbf{H}_{:,\mathcal{Q}_{j-1}}^H \mathbf{H}_{:,\mathcal{Q}_{j-1}})^{-\frac{1}{2}}, \overline{p}_{j+1}^{\frac{1}{2}} (\mathbf{H}_{:,\mathcal{Q}_{j+1}}^H \mathbf{H}_{:,\mathcal{Q}_{j+1}})^{-\frac{1}{2}}, \ldots, \overline{p}_r^{\frac{1}{2}} (\mathbf{H}_{:,\mathcal{Q}_r}^H \mathbf{H}_{:,\mathcal{Q}_r})^{-\frac{1}{2}})$.

From $E[\mathbf{Z}|\mathbf{G}_{k-1}]$ in (48), the global and exponential convergence performance of the conditional randomized iteration in ERIPA can be readily verified.

Theorem 3: Regarding the downlink massive MIMO, let $\mathcal{Q}_i$ be randomly sampled by the conditional sampling probability $\overline{p}_i$ in (47), ERIPA is convergent by

$$E[\|\mathbf{V}(\mathbf{t}^k - \mathbf{t}^*)\|^2] \leq \overline{\rho} \|\mathbf{V}(\mathbf{t}^{k-1} - \mathbf{t}^*)\|^2 \tag{49}$$

with a global and exponentially fast convergence rate

$$\overline{\rho} = 1 - \lambda_{\min}(\mathbf{H}^H \mathbf{H} E[\mathbf{Z}|\mathcal{Q}_j]) < 1. \tag{50}$$

Here, the Theorem can be proved by following a similar approach as that for Theorem 2. Based on it, we have the following result to illustrate the introduced convergence gain.

Corollary 2: With $\overline{p}_i$ in (47), ERIPA achieves an improved convergence than that with p_i in (18) due to a smaller convergence upper bound.

Proof: With the conditional sampling probability $\overline{p}_i$, the convergence rate of the randomized iteration is upper bounded by

$$\begin{aligned} \overline{\rho} &= 1 - \lambda_{\min}(\mathbf{H}^H \mathbf{H} E[\mathbf{Z}|\mathcal{Q}_j]) \\ &\leq 1 - \lambda_{\min}(\mathbf{H}^H \mathbf{H}) \lambda_{\min}(\overline{\mathbf{J}}^2) \\ &\leq 1 - \min_i \left\{ \overline{p}_i \cdot \frac{\lambda_{\min}(\mathbf{H}^H \mathbf{H})}{\lambda_{\max}(\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i} | i \neq j)} \right\}. \end{aligned} \tag{51}$$

As a comparison, with sampling probability p_i, the convergence rate is upper bounded by

$$\begin{aligned} \rho = 1 - \lambda_{\min}(\mathbf{H}^H \mathbf{H} E[\mathbf{Z}]) < 1 \\ - \min_i \left\{ p_i \cdot \frac{\lambda_{\min}(\mathbf{H}^H \mathbf{H})}{\lambda_{\max}(\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i})} \right\}. \end{aligned} \tag{52}$$

As a result, one can observe that $\overline{\rho}$ attains a smaller upper bound than ρ by

$$\min_i \left\{ \frac{\overline{p}_i}{\lambda_{\max}(\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i} | i \neq j)} \right\} > \min_i \left\{ \frac{p_i}{\lambda_{\max}(\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i})} \right\}, \tag{53}$$

due to the facts $\lambda_{\max}(\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i} | i \neq j) \leq \lambda_{\max}(\mathbf{H}_{:,\mathcal{Q}_i}^H \mathbf{H}_{:,\mathcal{Q}_i})$ and $\overline{p}_i > p_i$. ■

B. Upgrade by Multi-Step Conditional Sampling

By incorporating conditional sampling, the convergence performance of the algorithm can be improved by considering more sampling results from previous iterations. Typically, this corresponds to the following multi-step conditional sampling probability

$$\overline{p}_i^f \triangleq \mathcal{D}(\mathcal{Q}_i | \mathcal{Q}_j, \ldots, \mathcal{Q}_l) \tag{54}$$

with $\mathcal{Q}_i \notin \{\mathcal{Q}_j, \ldots, \mathcal{Q}_l\}$, where $\mathcal{Q}_l$ denotes the sampling choice of the $(k-f)$-th randomized iteration. In other words, all the previous f randomized iterations, i.e., $k-f, \ldots, k-1$, are considered in sampling of $\mathcal{Q}_i$ at the current k-th iteration, where the conditional sampling probability $\overline{p}_i$ in (47) is actually a special version of $\overline{p}_i^f$ when $f = 1$. Moreover, the sampling probability p_i in (18) is essentially the same as $\overline{p}_i^f$ with $f = 0$. Based on the multi-step conditional sampling probability $\overline{p}_i^f$, we have the Theorem below, where the proof is omitted due to brevity.

Theorem 4: Regarding the downlink massive MIMO, let $\mathcal{Q}_i$ be sampled randomly from the conditional sampling probability $\overline{p}_i^f$ in (54), ERIPA is convergent by

$$E[\|\mathbf{V}(\mathbf{t}^k - \mathbf{t}^*)\|^2] \leq \overline{\rho}^f \|\mathbf{V}(\mathbf{t}^{k-1} - \mathbf{t}^*)\|^2 \tag{55}$$

with a global and exponentially fast convergence rate

$$\overline{\rho}^f = 1 - \lambda_{\min}(\mathbf{H}^H \mathbf{H} E[\mathbf{Z}|\mathcal{Q}_j, \ldots, \mathcal{Q}_l]) < 1. \tag{56}$$

According to the above Theorem, the convergence gain introduced by the multi-step conditional sampling can be easily confirmed in the following, where the proof is similar to that in Corollary 2.

Corollary 3: As the increase of $1 \leq f \leq r-1$, the convergence of ERIPA with $\overline{p}_i^f$ in (54) gradually improves by a smaller convergence upper bound.

By Corollary 3, the smallest convergence upper bound is attained with $f = r-1$. More interestingly, with $f = r-1$, there is only one sampling option for $\mathcal{Q}_i$ if $k > r-1$. This means that the underlying sampling operation becomes deterministic. Clearly, this is beneficial to the implementation of ERIPA, which not only avoids the use of random sampling but also boosts the convergence performance. Therefore, to achieve more gains in both efficiency and convergence, multi-step conditional sampling based on $f = r-1$ is preferred in ERIPA, which substitutes p_i in (18) at step 3 of Alg. 1 by $\overline{p}_i^f$ in (54). As for future work, the efficient statistical methods can be further extended to hybrid precoding for jointly designing the baseband precoder and radio frequency (RF) precoder [52].

VI. Numerical Studies

In this section, the proposed ERIPA for downlink systems in massive MIMO is evaluated by simulations. In particular, the channel coefficients in matrix $\mathbf{H} \in \mathbb{C}^{N \times K}$ are assumed to be Gaussian distributed (i.e., corresponds to the Rayleigh channel), which is known perfectly at the BS. Meanwhile, following the works in [13], [14], and [38], the bit error rate (BER) with respect to the transmitted source $\mathbf{s}$ is applied to serve for the precoding performance comparison. This comes

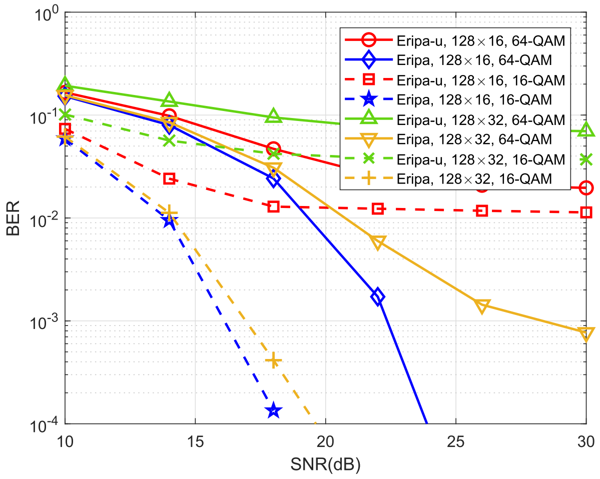


Fig. 2. BER versus the average SNR for 128 × 16 and 128 × 32 massive MIMO systems with 64-QAM and 16-QAM.

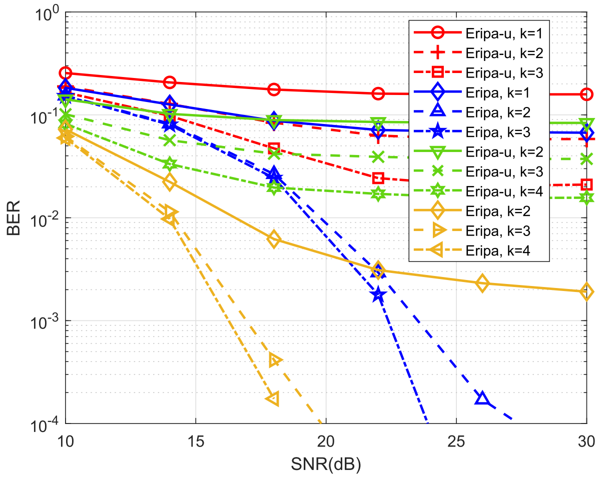


Fig. 3. BER versus the average SNR for a 128 × 16 massive MIMO system with 64-QAM and a 128 × 32 massive MIMO system with 16-QAM.

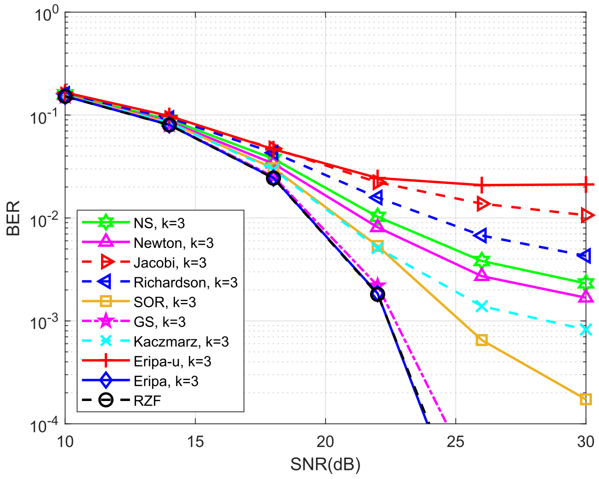


Fig. 4. BER versus the average SNR for a 128 × 16 massive MIMO system with 64-QAM.

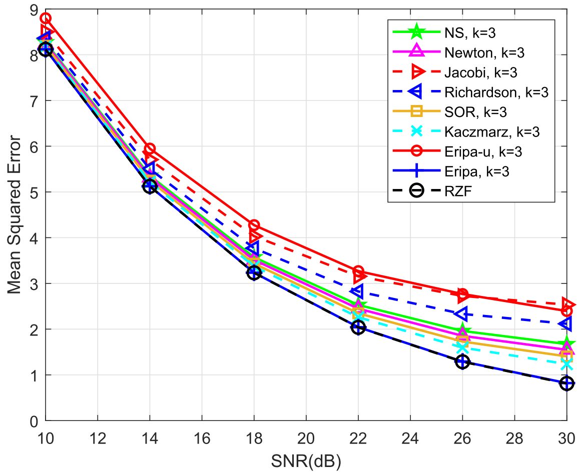


Fig. 5. Mean squared error (MSE) versus the average SNR for a 128 × 16 massive MIMO system with 64-QAM.

from the fact that the comparison between the estimated $\mathbf{t}$ and the linear solution $\mathbf{t}^*$ can be simply evaluated by the comparison between $\mathbf{s}$ and $\mathbf{s}^*$ due to the relationship $\mathbf{t} = \mathbf{A}^{-1}\mathbf{s}$. Here, we use "Eripa-u" and "Eripa" to respectively denote ERIPA with random sampling probability shown in (18) and with $r-1$ step conditional sampling probability shown in (54). Unless otherwise stated, the block size in ERIPA is set as $q = 8$ throughout the context. On the other hand, for a fair comparison with other iterative precoding algorithms, the full iteration with respect to ERIPA is adopted, which means that all the entries of $\mathbf{t}$ are updated in each iteration k.

In Fig. 2, the BER performance of ERIPA with $k = 3$ is illustrated in 128 × 16 and 128 × 32 massive MIMO systems with 64-QAM and 16-QAM, respectively. Clearly, in all cases, ERIPA with multi-step conditional sampling probability achieves a better precoding performance than that with random sampling probability. Undoubtedly, this is in accordance with the result given in Corollary 3. Meanwhile, one can observe that compared to the performance of ERIPA in 128 × 16, certain performance degradation is experienced with respect to the case 128 × 32. This can be easily interpreted as less received diversity gain is exploited due to the increment of the antennas at the user side given a fixed number of received antennas. Moreover, the change of the modulation size Q also has an impact upon the precoding performance, and a smaller modulation order naturally corresponds to a better BER performance.

In Fig. 3, under different numbers of full iterations, the BER performance of ERIPA is examined in a 128 × 16 massive MIMO system with 64-QAM (for $k = 1, 2, 3$) and a 128 × 32 massive MIMO system with 16-QAM (for $k = 2, 3, 4$). As expected, for both cases, the performance of ERIPA gradually improves with the increase of the full iteration number k, which are in line with the results described in Theorems 2 and 4. Undoubtedly, under the same number of full iterations, ERIPA with multi-step conditional sampling probability always obtains a better performance than that with the random sampling probability. Besides the faster convergence rate, the implementation of ERIPA with multi-step conditional sampling probability is also much more efficient, making it appealing for low-complexity linear precoding in downlink massive MIMO systems.

In Fig. 4, the BER performance comparison over various low-complexity iterative linear precoding schemes is illustrated in a 128 × 16 massive MIMO system with 64-QAM. Typically, the NS precoding in [22], the Newton precoding

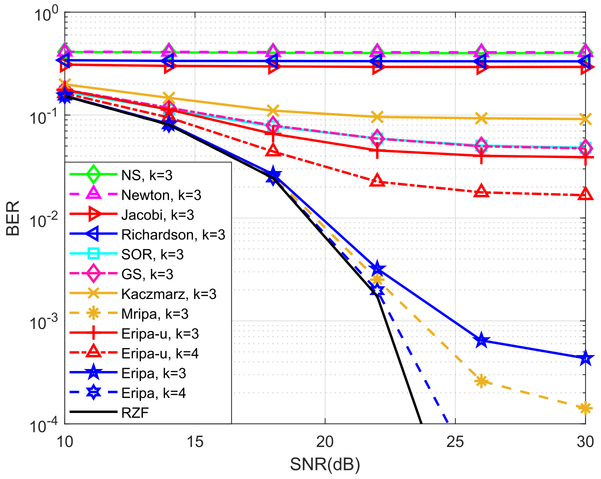


Fig. 6. BER versus the average SNR for a 128×16 massive MIMO system with 64-QAM with the normalized correlation coefficient $\psi = 0.05$.

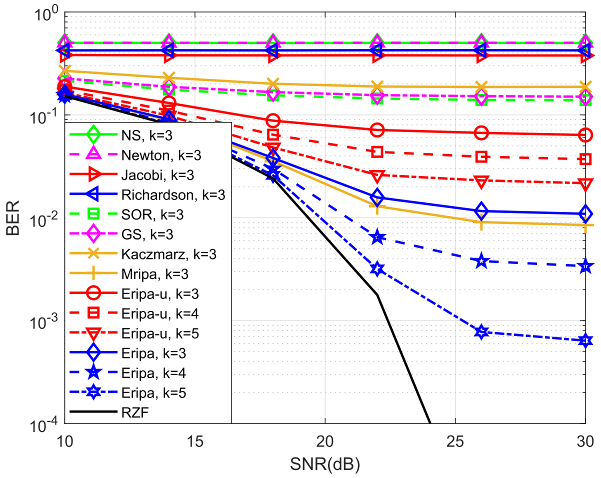


Fig. 7. BER versus the average SNR for a 128×16 massive MIMO system with 64-QAM with the normalized correlation coefficient $\psi = 0.1$.

in [25], the Jacobi precoding in [20], the Richardson precoding in [26] with the relaxation coefficient $\omega = 1/(N+K)$, the GS precoding in [28], the SOR precoding in [31] with the relaxation coefficient $\omega = \frac{2}{1+\sqrt{1-[\varrho(\mathbf{I}-\mathbf{D}^{-1}\mathbf{A})]^2}}$,[7] the Kaczmarz precoding in [39] are applied for comparison. As can be seen clearly, under the same iteration number $k = 3$, the proposed ERIPA achieves a competitive BER performance among them as a benefit of the fast convergence. More importantly, besides the convergence gain, ERIPA also entails a much lower complexity cost, which seamlessly incorporates both the matrix multiplication and inversion into the designed iterations. More specifically, near RZF precoding performance can be attained by ERIPA with $k = 3$. As a counterpart of Fig. 4, Fig. 5 illustrates the mean squared error (MSE) between the transmitted source information $\mathbf{s}$ and the received information $\widehat{\mathbf{s}}$ at the user terminal (i.e., $\|\mathbf{s} - \widehat{\mathbf{s}}\|$). Intuitively, according to (6), a better estimation of precoding matrix $\mathbf{G}$ or vector $\mathbf{t}$ naturally correspond to a better precoding performance, so as to a smaller MSE $\|\mathbf{s} - \widehat{\mathbf{s}}\|$. As clearly can be observed, under the number of full iteration $k = 3$, the MSE of all the precoding schemes decrease gradually with the improvement of SNR. More specifically, the proposed ERIPA converges in a superior speed than other schemes and near RZF performance in terms of MSE can be easily achieved by it, which are in accordance with the results shown in Fig. 4.

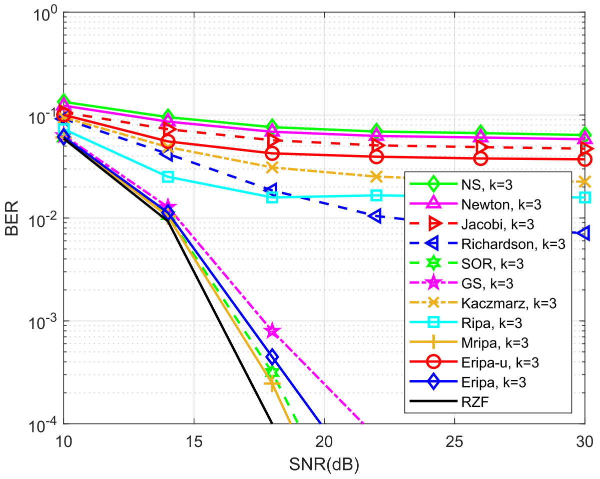


Fig. 8. BER versus the average SNR for a 128×32 massive MIMO system with 16-QAM.

Other than independent, identically, distributed (i.i.d.) Gaussian fading channels, the impact of correlated channels for the proposed ERIPA is also investigated in Fig. 6, where a 128×16 massive MIMO system with 64-QAM is considered. To be more specific, the correlated channel matrix $\mathbf{R}_\text{b}^{\frac{1}{2}}\mathbf{H}\mathbf{R}_\text{u}^{\frac{1}{2}}$ in [53] is applied. $\mathbf{R}_\text{b} \in \mathbb{C}^{N\times N}$ and $\mathbf{R}_\text{u} \in \mathbb{C}^{K\times K}$ indicate the correlation matrices at the transmitter and receiver sides respectively, i.e.,

$$\mathbf{R}_\text{b} = \begin{bmatrix} 1 & \psi & \psi^4 & \cdots & \psi^{(N-1)^2} \\ \psi & 1 & \psi & \cdots & \vdots \\ \psi^4 & \psi & 1 & \cdots & \psi^4 \\ \vdots & \vdots & \vdots & \ddots & \psi \\ \psi^{(N-1)^2} & \cdots & \psi^4 & \psi & 1 \end{bmatrix}$$

and $\mathbf{R}_\text{u}$ obeys the same structure as $\mathbf{R}_\text{b}$ but with K instead of N. Meanwhile, the normalized correlation coefficient $1 \geq \psi \geq 0$ is exploited to adjust the underlying correlation, where $\psi = 0$ corresponds to an uncorrelated scenario and vice versa. Compared to the results for i.i.d. channels in Fig. 4, the performance of all the iterative linear precoding schemes deteriorate accordingly in Fig. 6 with $\psi = 0.05$. This can be well explained as a more correlated channel means a higher condition number, which naturally results in a slower convergence performance. Interestingly, due to the correlated channels, the precoding schemes based on NS, Newton, Jacobi, Richardson and so on do not perform well at all. In contrast, the proposed ERIPA still works as the benefit of global convergence. Note that the precoding schemes RIPA and MRIPA from [38] are also added in Fig. 6 for a better comparison. Although MRIPA enjoys a slightly better BER performance than ERIPA, it requires a more higher complexity cost than ERIPA and more details can be found in Table I.

[7] $\varrho(\cdot)$ stands for the matrix spectral radius.

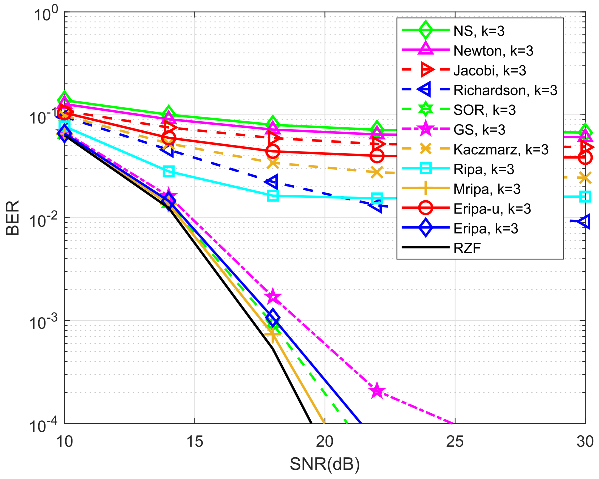


Fig. 9. BER versus the average SNR for a 128×32 massive MIMO system with 16-QAM under imperfect CSI.

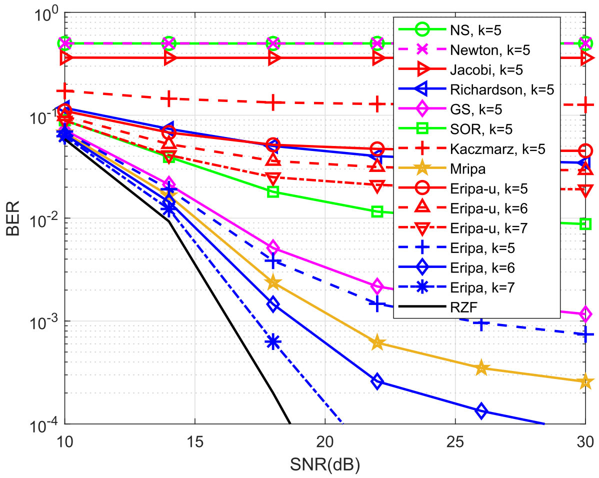


Fig. 10. BER versus the average SNR for a 128×64 massive MIMO system with 16-QAM.

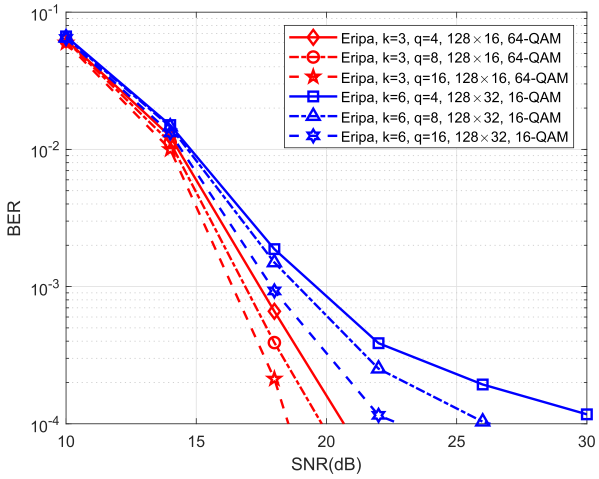


Fig. 11. BER versus the average SNR for a 128×16 massive MIMO system with 64-QAM and a 128×32 massive MIMO system with 16-QAM.

To be specific, Fig. 7 is given but with an improved normalized correlation coefficient $\psi = 0.1$. Clearly, the performance of all the precoding schemes are degraded accordingly. Nevertheless, near RZF performance can still be achieved by ERIPA with a larger size k.

In Fig. 8, the BER performance comparison over different low-complexity iterative linear precoding schemes is shown for a 128×32 massive MIMO system with 16-QAM. Compared to the case of 128×16, the received diversity gain in a 128×32 massive MIMO system is reduced due to the increased number of antennas at the user side. As such, the advantage like favorable propagation can not be well exploited, and the other precoding algorithms like NS, Newton, Jacobi and so on fail to work. On the contrary, the proposed ERIPA still works well to offer the competitive BER performance but with the reduced complexity cost. On the other hand, the channel with imperfect channel state information (CSI) is studied as well, and Fig. 9 is illustrated as a counterpart of Fig. 8. In particular, by letting the channel matrix with imperfect CSI as $\widehat{\mathbf{H}} = \mathbf{H} + \Delta\mathbf{H}$, the case of imperfect CSI is taken into account in the linear precoding, where $\Delta\mathbf{H}$ is the channel estimation error. Here, we set each of its elements follows $\mathcal{CN}(0, \sigma_e^2)$ with $\sigma_e^2 = 0.1$ [54], [55]. Intuitively, compared to the perfect case in Fig. 8, all the performance of the linear precoding schemes (including RZF) degrade accordingly in Fig. 9. Nevertheless, the near RZF performance still can be attained by the proposed ERIPA with a moderate iteration number k.

In Fig. 10, the BER performance comparison over various iterative precoding schemes in a 128×64 massive MIMO system with 16-QAM is provided. Clearly, in this case, only a few iterative precoding schemes with global convergence can work, and ERIPA has a competitive precoding performance among them. On the other hand, compared to the cases of 128×16 and 128×32, more iterations are required to approach the near NZF performance. This is easy to understand as less received diversity gain can be utilized. Nevertheless, the precoding performance still improves with the increment of iterations. Note that although MRIPA has a better performance than the proposed ERIPA under the same iterations, its computational complexity is indeed much higher than that of ERIPA.

In Fig. 11, the various choices of q in the proposed ERIPA are studied respectively in a 128×16 massive MIMO system with 64-QAM and a 128×32 massive MIMO system with 16-QAM. Typically, in both two cases, better precoding performance can be achieved with the increase of $q = 4, 8, 16$. This is because a larger size q means more elements of $\mathbf{t}$ are processed together, where the correlation among them can be beneficially utilized therein. However, as shown in (17) and (19), a larger size q also incurs higher computational complexity cost, implying a non-trivial trade-off between the performance and complexity. As such, the block size q should be carefully selected for satisfying the various requirements.

In Fig. 12, the complexity compassion over iterative linear precoding schemes is presented in terms of the flops with respect to a full iteration. To be more specific, Table II is also given for a better understanding. As expected, with the increment of K, all the complexity costs of iterative precoding schemes improve gradually. Among them, the precoding schemes based NS and Newton methods require much higher number of flops compared with others. This is because the iterations in them are performed with respect to the matrices rather than vectors. Meanwhile, the precoding schemes based

TABLE II

AVERAGE COMPLEXITY COMPARISON IN FLOPS WITH RESPECT TO A FULL ITERATION IN VARIOUS ITERATIVE LINEAR PRECODING SCHEMES FOR A 128 × K MASSIVE MIMO SYSTEM

	$K = 16$	$K = 24$	$K = 32$	$K = 40$	$K = 48$	$K = 56$	$K = 64$
NS	66528	186576	397248	723120	1188768	1818768	2637696
Newton	47072	124368	253888	447920	718752	1078672	1539968
Jacobi	35840	80640	143360	224000	322560	439040	573440
SOR	36032	80928	143744	224480	323136	439712	574208
MRIPA	36848	82793	147070	229680	330610	449870	587456
ERIPA	37824	65661	97701	133425	172473	214593	259584

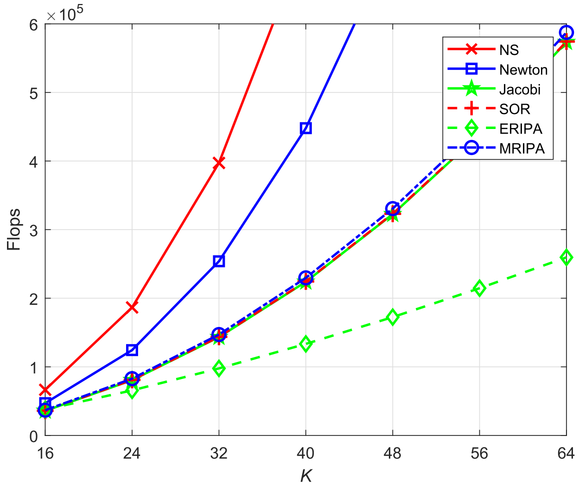


Fig. 12. Complexity comparison in terms of flops for a 128 × K massive MIMO system.

on Jacobi, SOR and MRIPA incur the same level complexity cost. Different from them, by designing the iterations simplifying both of matrix multiplication and inversion, the proposed ERIPA achieves a much lower complexity. Typically, with $K = 40$, about 37% percent computational complexity can be saved by ERIPA compared with MRIPA. More importantly, as shown in Table I, with the increment of the system dimension, a substantially higher order of computational complexity can be further reduced. For example, as for the case of $K = 64$, more than 50% percent complexity can be reduced by ERIPA. Therefore, considerable complexity reduction with negligible performance loss can be obtained by ERIPA such that it is preferable in downlink massive MIMO systems.

VII. CONCLUSION

In this paper, a statistical low-complexity linear precoding, ERIPA, was proposed for computational complexity reduction, which performs the iterations by jointly considering the required matrix multiplication and inversion operations. By doing this, the overall computational complexity of linear precoding can be significantly reduced. Besides the complexity reduction, according to theoretical demonstrations, we showed that the proposed ERIPA not only enjoys an exponential convergence performance but also attains the global convergence. Meanwhile, the impacts of different random sampling probability on the convergence were also investigated, which give an explicit convergence rate for guiding the iterative operations.

On the other hand, by fully utilizing conditional sampling, further enhancements with respect to ERIPA were provided such that extra gains in terms of convergence and efficiency are obtained as well. Motivated by this, multi-step conditional sampling was adopted to ERIPA to facilitate its efficient implementation. Compared to the existing low-complexity iterative linear precoding schemes, ERIPA achieves a faster and global convergence performance with less complexity cost, thus making it rather promising in various practical scenarios of massive MIMO in practice. Finally, simulations were given to clearly illustrate the achieved gains of the proposed ERIPA in terms of both performance and complexity over different massive MIMO scenarios.

REFERENCES

[1] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020.

[2] I. Tomkos, D. Klonidis, E. Pikasis, and S. Theodoridis, "Toward the 6G network Era: Opportunities and challenges," *IT Prof.*, vol. 22, no. 1, pp. 34–38, Jan. 2020.

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

[4] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[5] C. Ding, J.-B. Wang, H. Zhang, M. Lin, and J. Wang, "Joint MU-MIMO precoding and resource allocation for mobile-edge computing," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1639–1654, Mar. 2021.

[6] Y. Wu, C.-K. Wen, C. Xiao, X. Gao, and R. Schober, "Linear precoding for the MIMO multiple access channel with finite alphabet inputs and statistical CSI," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 983–997, Feb. 2015.

[7] Z. Qiu, S. Zhou, M. Zhao, and W. Zhou, "Low-complexity precoding by exploiting spatial sparsity in massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 4740–4753, Jul. 2022.

[8] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local partial zero-forcing precoding for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4758–4774, Jul. 2020.

[9] E. Shi, J. Zhang, J. Zhang, D. W. K. Ng, and B. Ai, "Decentralized coordinated precoding design in cell-free massive MIMO systems for URLLC," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2638–2642, Feb. 2023.

[10] J. Hoydis, S. T. Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[11] F. Rusek et al., "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[12] N. Fatema, G. Hua, Y. Xiang, D. Peng, and I. Natgunanathan, "Massive MIMO linear precoding: A survey," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3920–3931, Dec. 2018.

[13] Y. Liu, J. Liu, Q. Wu, Y. Zhang, and M. Jin, "A near-optimal iterative linear precoding with low complexity for massive MIMO systems," *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 1105–1108, Jun. 2019.

[14] T. Xie, L. Dai, X. Gao, X. Dai, and Y. Zhao, "Low-complexity SSOR-based precoding for massive MIMO systems," *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 744–747, Apr. 2016.

[15] 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 Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1069–1084, Feb. 2019.

[16] A. Kammoun, A. Müller, E. Björnson, and M. Debbah, "Linear precoding based on polynomial expansion: Large-scale multi-cell MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 861–875, Oct. 2014.

[17] J.-C. Chen, C.-J. Wang, K.-K. Wong, and C.-K. Wen, "Low-complexity precoding design for massive multiuser MIMO systems using approximate message passing," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5707–5714, Jul. 2016.

[18] C. Zhang, Y. Jing, Y. Huang, and L. Yang, "Performance analysis for massive MIMO downlink with low complexity approximate zero-forcing precoding," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3848–3864, Sep. 2018.

[19] C. Jeon, Z. Li, and C. Studer, "Approximate gram-matrix interpolation for wideband massive MU-MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 4677–4688, May 2020.

[20] J. Minango and C. de Almeida, "A low-complexity linear precoding algorithm based on Jacobi method for massive MIMO systems," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–5.

[21] D. Zhu, B. Li, and P. Liang, "On the matrix inversion approximation based on Neumann series in massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 1763–1769.

[22] B. Nagy, M. Elsabrouty, and S. Elramly, "Fast converging weighted Neumann series precoding for massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 154–157, Apr. 2018.

[23] H. V. Nguyen, V.-D. Nguyen, and O.-S. Shin, "Low-complexity precoding for sum rate maximization in downlink massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 2, pp. 186–189, Apr. 2017.

[24] Y. Saad, *Iterative Methods for Sparse Linear Systems*. Philadelphia, PA, USA: SIAM, 1997.

[25] Y. Man, C. Zhang, Z. Li, F. Yan, S. Xing, and L. Shen, "Massive MIMO pre-coding algorithm based on improved Newton iteration," in *Proc. IEEE 85th Veh. Technol. Conf. (VTC Spring)*, Jun. 2017, pp. 1–5.

[26] Z. Lu, J. Ning, Y. Zhang, T. Xie, and W. Shen, "Richardson method based linear precoding with low complexity for massive MIMO systems," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–4.

[27] A. Björck, *Numerical Methods for Least Squares Problems*. Philadelphia, PA, USA: SIAM, 1996.

[28] M. A. Albreem, A. H. Al Habbash, A. M. Abu-Hundrouss, and S. S. Ikki, "Overview of precoding techniques for massive MIMO," *IEEE Access*, vol. 9, pp. 60764–60801, 2021.

[29] A. Greenbaum, *Iterative Methods for Solving Linear Systems*. Philadelphia, PA, USA: SIAM, 1997.

[30] X. Gao, L. Dai, J. Zhang, S. Han, and I. Chih-Lin, "Capacity-approaching linear precoding with low-complexity for large-scale MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 1577–1582.

[31] T. Xie, Q. Han, H. Xu, Z. Qi, and W. Shen, "A low-complexity linear precoding scheme based on SOR method for massive MIMO systems," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–5.

[32] S. Zarei, W. Gerstacker, and R. Schober, "Low-complexity widely-linear precoding for downlink large-scale MU-MISO systems," *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 665–668, Apr. 2015.

[33] Y. Man, Z. Li, F. Yan, S. Xing, and L. Shen, "Massive MIMO pre-coding algorithm based on truncated Kapteyn series expansion," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Dec. 2016, pp. 1–5.

[34] P. Yang and H. Yang, "A low-complexity linear precoding for MIMO channels with finite constellation inputs," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1415–1418, Oct. 2019.

[35] D. Kwon, W.-Y. Yeo, and D. K. Kim, "A new precoding scheme for constructive superposition of interfering signals in multiuser MIMO systems," *IEEE Commun. Lett.*, vol. 18, no. 11, pp. 2047–2050, Nov. 2014.

[36] S. Zarei, W. H. Gerstacker, R. Weigel, M. Vossiek, and R. Schober, "Robust MSE-balancing hierarchical linear/Tomlinson–Harashima precoding for downlink massive MU-MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7309–7324, Nov. 2018.

[37] W. Zhang et al., "Widely linear precoding for large-scale MIMO with IQI: Algorithms and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3298–3312, May 2017.

[38] Z. Wang, R. M. Gower, C. Zhang, S. Lyu, Y. Xia, and Y. Huang, "A statistical linear precoding scheme based on random iterative method for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10115–10129, Dec. 2022.

[39] M. N. Boroujerdi, S. Haghighatshoar, and G. Caire, "Low-complexity statistically robust precoder/detector computation for massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6516–6530, Oct. 2018.

[40] V. Croisfelt, A. Amiri, T. Abr ao, E. De Carvalho, and P. Popovski, "Accelerated randomized methods for receiver design in extra-large scale MIMO arrays," *IEEE Trans. Veh. Technol.*, vol. 70, no. 7, pp. 6788–6799, Jul. 2021.

[41] P. Zuo, Z. Sun, and R. Huang, "Extremely-fast, energy-efficient massive MIMO precoding with analog RRAM matrix computing," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 70, no. 7, pp. 2335–2339, Jul. 2023.

[42] M. M. Mojahedian, R. Mosayebi, and A. Lozano, "Pseudo-inverse vs generalized inverse for C-RAN downlink precoding," in *Proc. IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[43] R. W. Heath and A. Lozano, *Foundations of MIMO Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[44] H. Prabhu, J. Rodrigues, O. Edfors, and F. Rusek, "Approximative matrix inverse computations for very-large MIMO and applications to linear pre-coding systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2013, pp. 2710–2715.

[45] A. Müller, A. Kammoun, E. Björnson, and M. Debbah, "Efficient linear precoding for massive MIMO systems using truncated polynomial expansion," in *Proc. IEEE 8th Sensor Array Multichannel Signal Process. Workshop (SAM)*, Jul. 2014, pp. 273–276.

[46] F. Rosário, F. A. Monteiro, and A. Rodrigues, "Fast matrix inversion updates for massive MIMO detection and precoding," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 75–79, Jan. 2016.

[47] A. Björck, *Numerical Methods in Matrix Computations*. Switzerland: Springer, 2015.

[48] F. A. Monteiro and I. J. Wassell, "Recovery of a lattice generator matrix from its Gram matrix for feedback and precoding in MIMO," in *Proc. 4th Int. Symp. Commun., Control Signal Process. (ISCCSP)*, Mar. 2010, pp. 1–6.

[49] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 4th ed. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 2013.

[50] E. Anarakifirooz and S. Loyka, "Favorable propagation for massive MIMO with circular and cylindrical antenna arrays," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 458–462, Mar. 2022.

[51] T. V. Chien, H. Q. Ngo, S. Chatzinotas, and B. Ottersten, "Reconfigurable intelligent surface-assisted massive MIMO: Favorable propagation, channel hardening, and rank deficiency [lecture notes]," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 97–104, May 2022.

[52] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, Dec. 2014.

[53] H. R. Bahrami, T. Le-Ngoc, A. M. N. Nasrabadi, and S. H. Jamali, "Precoder design based on correlation matrices for MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 3, May 2005, pp. 2001–2005.

[54] D. L. Colon, F. H. Gregorio, and J. Cousseau, "Linear precoding in multi-user massive MIMO systems with imperfect channel state information," in *Proc. 16th Workshop Inf. Process. Control (RPIC)*, Oct. 2015, pp. 1–6.

[55] N. Lee, O. Simeone, and J. Kang, "The effect of imperfect channel knowledge on a MIMO system with interference," *IEEE Trans. Commun.*, vol. 60, no. 8, pp. 2221–2229, Aug. 2012.

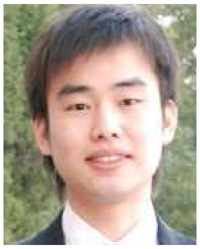
Zheng Wang (Senior Member, IEEE) received the B.S. degree in electronic and information engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2009, the M.S. degree in communications from The University of Manchester, Manchester, U.K., in 2010, and the Ph.D. degree in communication engineering from the Imperial College London, U.K., in 2015.

Since 2021, he has been an Associate Professor with the School of Information and Engineering, Southeast University, Nanjing. From 2015 to 2016, he was a Research Associate with the Imperial College London. From 2016 to 2017, he was an Senior Engineer with the Radio Access Network R&D Division, Huawei Technologies Company. From 2017 to 2020, he was an Associate Professor with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing. Since 2023, he has been recognized as a Highly Cited Chinese Researcher by Elsevier for exceptional research performance in the field of information and communications engineering. His current research interests include massive MIMO systems, machine learning and data analytics over wireless networks, and lattice theory for wireless communications. He received the Spark Award from Huawei Technologies Company in 2023 for the research works toward spark challenge about efficient baseband matrix processing. He is also a Youth Editor of *Chinese Journal of Electronics* (CJE).

Le Liang (Member, IEEE) received the B.E. degree in information engineering from Southeast University, Nanjing, China, in 2012, the M.A.Sc. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 2015, and the Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2018. From 2019 to 2021, he was a Research Scientist with the Intel Laboratories, Hillsboro, OR, USA. Since 2021, he has been with the National Mobile Communications Research Laboratory, Southeast University. His main research interests are in wireless communications, signal processing, and machine learning.

He is a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society. He received the Best Paper Award of IEEE/CIC ICCC in 2014. He serves as an Associate Editor for IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING and *China Communications*. He was an Associate Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (IEEE JSAC) Series on Machine Learning in Communications and Networks from 2020 to 2022 and an Editor for the IEEE COMMUNICATIONS LETTERS from 2019 to 2023.

Shanxiang Lyu received the B.S. and M.S. degrees in electronic and information engineering from the South China University of Technology, Guangzhou, China, in 2011 and 2014, respectively, and the Ph.D. degree from the Electrical and Electronic Engineering Department, Imperial College London, U.K., in 2018. He is currently an Associate Professor with the College of Cyber Security, Jinan University, Guangzhou, China. His research interests include quantization theory and information security. He was a recipient of the 2021 CIE Information Theory Society Yong-Star Award and the 2020 Superstar Supervisor Award of the National Crypto-Math Challenge of China. He was in the Organizing Committee of Inscrypt 2020.



Yili Xia (Member, IEEE) received the B.Eng. degree in information engineering from Southeast University, Nanjing, China, in 2006, the M.Sc. degree (Hons.) in communications and signal processing from the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., in 2007, and the Ph.D. degree in adaptive signal processing from the Imperial College London in 2011.

Since 2013, he has been an Associate Professor in signal processing with the School of Information Science and Engineering, Southeast University, where he is currently the Deputy Head of the Department of Information and Signal Processing Engineering. His research interests include complex and hyper-complex statistical analysis, detection and estimation, linear and nonlinear adaptive filters, and their applications on communications and power systems. He was a recipient of the Best Student Paper Award at the International Symposium on Neural Networks (ISNN) in 2010 (coauthor) and the Education Innovation Award at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2019. He is currently an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Yongming Huang (Senior Member, IEEE) received the B.S. and M.S. degrees from Nanjing University, Nanjing, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from Southeast University, Nanjing, in 2007.

Since March 2007, he has been a Faculty Member of the School of Information Science and Engineering, Southeast University, where he is currently a Full Professor. From 2008 to 2009, he visited the Signal Processing Laboratory, Royal Institute of Technology, Stockholm, Sweden. He has authored or coauthored more than 200 peer-reviewed articles and holds more than 80 invention patents. His research interests include intelligent 5G/6G mobile communications and millimeter wave wireless communications. He submitted around 20 technical contributions to IEEE standards and was awarded a certificate of appreciation for outstanding contribution to the development of IEEE standard 802.11aj. He was an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING and a Guest Editor of the IEEE JOURNAL SELECTED AREAS IN COMMUNICATIONS. He is currently an Editor-at-Large of the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY and an Associate Editor of IEEE WIRELESS COMMUNICATIONS LETTERS.

Derrick Wing Kwan Ng (Fellow, IEEE) received the bachelor's (Hons.) and Master of Philosophy degrees in electronic engineering from The Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2006 and 2008, respectively, and the Ph.D. degree from The University of British Columbia, Vancouver, BC, Canada, in November 2012.

Following the Ph.D. degree, he was a Senior Post-Doctoral Fellow with the Institute for Digital Communications. He has been recognized as a Highly Cited Researcher by Clarivate Analytics (Web of Science) since 2018. He was a recipient of Australian Research Council (ARC) Discovery Early Career Researcher Award in 2017, IEEE Communications Society Leonard G. Abraham Prize in 2023, IEEE Communications Society Stephen O. Rice Prize in 2022, Best Paper Awards at the WCSP in 2020 and 2021, IEEE TCGCC Best Journal Paper Award in 2018, INISCOM in 2018, IEEE International Conference on Communications (ICC) in 2018, 2021, 2023, and 2024, IEEE International Conference on Computing, Networking and Communications (ICNC) in 2016, IEEE Wireless Communications and Networking Conference (WCNC) in 2012, IEEE Global Telecommunication Conference (Globecom) in 2011, 2021, and 2023, and IEEE Third International Conference on Communications and Networking in China in 2008. From January 2012 to December 2019, he served as an Editorial Assistant to the Editor-in-Chief for IEEE TRANSACTIONS ON COMMUNICATIONS. He is also an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS and the Associate Editor-in-Chief for the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.