Capacity-Approaching Linear Precoding with Low-Complexity for Large-Scale MIMO Systems

Xinyu Gao[1], Linglong Dai[1], Jiayi Zhang[1], Shuangfeng Han[2], and Chih-Lin I[2]
[1]Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Green Communication Research Center, China Mobile Research Institute, Beijing 100053, China
E-mail: daill@tsinghua.edu.cn, hanshuangfeng@chinamobile.com

***Abstract*—Linear precoding techniques, such as zero forcing precoding, can achieve the near-optimal capacity due to the favorable channel propagation in large-scale MIMO systems, but involve complicated matrix inversion of large size. In this paper, we propose a low-complexity linear precoding scheme based on the Gauss-Seidel (GS) method. The proposed scheme can achieve the capacity-approaching performance of the classical linear precoding schemes in an iterative way without complicated matrix inversion, which can reduce the overall complexity by one order of magnitude. We also prove that the proposed GS-based precoding scheme has a faster convergence rate than the recently proposed Neumann-based precoding scheme. Simulation results demonstrate that the proposed scheme can achieve the exact capacity-approaching performance of the classical linear precoding schemes with only a small number of iterations.**

I. Introduction

Multiple-input multiple-output (MIMO) technology has been successfully integrated in a series of well established communication technologies, such as the 4th generation (4G) cellular system LTE-A, IEEE 802.11n wireless LAN system, etc. It is considered as a promising key technology for future wireless systems [1], [2]. Unlike the traditional small-scale MIMO (e.g., at most 8 antennas in LTE-A), large-scale MIMO, which equips a very large number of antennas (e.g., 256 antennas or even more) at the base station (BS) to simultaneously serve multiple users, is recently proposed [3]. It has been theoretically proved that large-scale MIMO can achieve orders of simultaneous increase in spectrum and energy efficiency [4].

However, realizing the very attractive merits of large-scale MIMO in practice faces several challenging problems, one of which is the low-complexity precoding in the downlink [5]. In order to shift the complicated processing of multi-user interference cancellation from the users to the BS to relieve the computational complexity of users in the downlink, two categories of precoding techniques, i.e., nonlinear and linear precoding, have been proposed. The optimal nonlinear precoding technique is the dirty paper precoding (DPC), which has been proved to be able to achieve the ideal channel capacity by subtracting the potential interferences before transmission [6]. However, it is very difficult to be realized for large-scale MIMO systems due to the high complexity of successive encoding and decoding. To achieve the close-optimal capacity with reduced complexity, some other nonlinear precoding techniques, such as the vector perturbation (VP) precoding [7] and the lattice-aided precoding [8] have been proposed, but their complexity is still unaffordable when the dimension of the MIMO system is large or the modulation order is high [9] (e.g., 256 antennas at the BS with 64 QAM modulation). To make a trade-off between the capacity and complexity, one can resort to linear precoding techniques, which can also achieve the capacity-approaching performance, since the favorable channel propagation makes the channel matrix asymptotically orthogonal in large-scale MIMO systems [5]. The simplest linear precoding scheme is the match filter (MF) precoding, but it can only achieve the satisfying capacity when the number of antennas at BS is very large (e.g., 1024 or more), while for a more realistic large-scale MIMO system, the zero forcing (ZF) precoding can enjoy a much better performance than the MF precoding [5]. However, the ZF precoding involves unfavorable complicated matrix inversion whose complexity is cubic with respect to the number of users. Very recently, the ZF precoding based on Neumann series approximation (Neumann-based precoding) was proposed in [10] to reduce the computational complexity of matrix inversion, which is realized by converting the matrix inversion into a series of matrix-vector multiplications. However, only marginal complexity reduction can be achieved.

In this paper, we propose a low-complexity capacity-approaching linear precoding based on the Gauss-Seidel (GS) method [11]. The proposed GS-based precoding scheme can reconstruct the transmitted signal in an iterative way without the complicated matrix inversion, and the analysis shows that the overall complexity can be reduced by one order of magnitude. We also prove that the proposed GS-based precoding scheme has a faster convergence rate than the recently proposed Neumann-based precoding scheme [10]. Simulation results verify that the proposed precoding scheme can achieve a satisfying capacity in a small number of iterations. To the best of our knowledge, this work is the first one to utilize the GS method for the precoding in large-scale MIMO systems.

The rest of the paper is organized as follows. Section II briefly introduces the system model. Section III specifies the proposed low-complexity capacity-approaching precoding scheme, together with the convergence rate and the complexity analysis. The simulation results of the achieved channel capacity and the bit error rate (BER) performance are shown in Section IV. Finally, conclusions are drawn in Section V.

978-1-4673-6432-4/15/$31.00 ©2015 IEEE

Notation: Lower-case and upper-case boldface letters denote vectors and matrices, respectively; $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^{-1}$, $\det(\cdot)$, and $\mathrm{tr}(\cdot)$ denote the transpose, conjugate transpose, inversion, determinant, and trace of a matrix, respectively; $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the Frobenius norm of a matrix and the 2-norm of a vector, respectively; $|\cdot|$ and $(\cdot)^*$ denote the absolute and conjugate operators, respectively; $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denote the real part and imaginary part of a complex number, respectively; Finally, $\mathbf{I}_N$ is the $N \times N$ identity matrix.

II. System Model

As illustrated in Fig. 1, we consider a single cell multi-user large-scale MIMO system in the downlink, which employs N antennas at the BS to simultaneously serve K single-antenna users [3]–[5]. In such system, we usually have $N \gg K$, e.g., $N = 256$ and $K = 16$ were considered in [5]. The received signal vector $\mathbf{y} = [y_1, \cdots, y_K]^T$ containing the received signals for K users can be represented as

$$\mathbf{y} = \sqrt{\rho_f}\mathbf{H}\mathbf{t} + \mathbf{n}, \tag{1}$$

where ρ_f is the signal-to-noise ratio (SNR) in the downlink, $\mathbf{H} \in \mathbb{C}^{K\times N}$ denotes the flat Rayleigh fading channel matrix whose entries follow the distribution $\mathcal{CN}(0, 1)$, $\mathbf{n} = [n_1, \cdots, n_K]^T$ presents the additive white Gaussian noise (AWGN) vector with independent and identically distributed (i.i.d.) zero mean and unit-variance complex Gaussian random variables, $\mathbf{t}$ denotes the $N \times 1$ normalized signal vector for actual transmission after precoding (i.e., $\mathrm{E}\left\{\|\mathbf{t}\|^2\right\} = K$), which is obtained by

$$\mathbf{t} = \mathbf{P}\mathbf{s}, \tag{2}$$

where $\mathbf{P}$ is the $N \times K$ precoding matrix, $\mathbf{s} = [s_1, \cdots, s_K]^T$ presents the original signal vector for all K users to be transmitted.

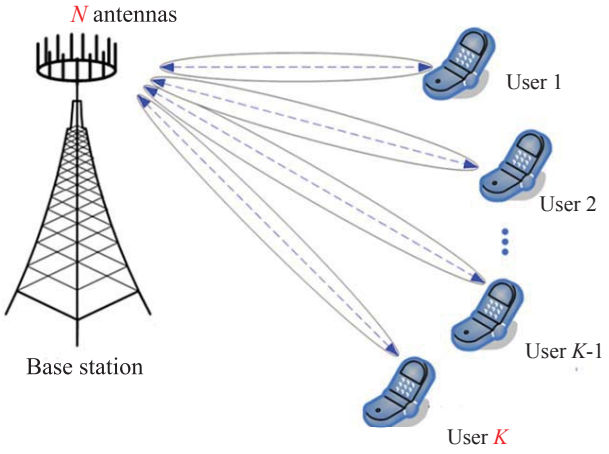

Fig. 1. Large-scale MIMO in the downlink.

The increased number of antennas N at the BS (while the number of users K is fixed) will lead to some special channel properties for large-scale MIMO. One attractive property is that the loss of signal power caused by imperfect channel state information (CSI) is less probable to induce the interference to other users [12], [13], which makes the linear precoding robust to the CSI mismatch. The other attractive property is the well-known favorable channel propagation, which makes the columns of channel matrix $\mathbf{H}$ asymptotically orthogonal due to the selected users are usually far away from each other [5]. That means the simple linear precoding techniques can achieve the capacity-approaching performance as will be verified later in Section IV.

III. Low-Complexity Linear Precoding Scheme For Large-Scale MIMO

In this section, we first briefly introduce the classical ZF precoding, which is capacity-approaching but involves high complexity for downlink large-scale MIMO systems. Then a low-complexity precoding scheme based on GS method is proposed to iteratively achieve the ZF precoding performance. The proof that the GS-based precoding scheme has faster convergence rate is also derived. Finally, we provide the complexity analysis of the proposed scheme to show its advantages over the recently proposed Neumann-based precoding scheme.

A. ZF precoding scheme

The ZF precoding is a scheme by which the multiple-antenna transmitter can eliminate multiuser interferences. The ZF precoding matrix can be presented as [5]

$$\mathbf{P}_{ZF} = \beta\mathbf{H}^{\dagger}, \tag{3}$$

where $\mathbf{H}^{\dagger} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1} = \mathbf{H}^H\mathbf{W}^{-1}$ denotes the pseudo-inversion of the channel matrix $\mathbf{H}$, here $\mathbf{W}$=$\mathbf{H}\mathbf{H}^H$, β is the normalized factor which averages the fluctuations in transmit power. A suitable choice for β is

$$\beta = \sqrt{\frac{K}{\mathrm{tr}\left(\mathbf{W}^{-1}\right)}}. \tag{4}$$

Then we can obtain the precoded signal vector $\mathbf{t}$ for transmission as

$$\mathbf{t} = \beta\mathbf{H}^H\mathbf{W}^{-1}\mathbf{s} = \beta\mathbf{H}^H\hat{\mathbf{s}}. \tag{5}$$

Considering (1) and (2), we can use $\mathbf{G} = \mathbf{H}\mathbf{P}$ to present the equivalent channel matrix. Since the CSI is assumed to be known at the transmitter [10], we have $|g_{mk}|^2 = 0$ if $m \neq k$, where g_{mk} is the element of $\mathbf{G}$ in the mth row and kth column. Then, the received signal to interference plus noise ratio (SINR) for any user k can be computed as

$$\begin{aligned}\gamma_k &= \frac{\frac{\rho_f}{K}|g_{kk}|^2}{\frac{\rho_f}{K}\sum_{m\neq k}^{K}|g_{mk}|^2+1} = \frac{\rho_f}{K}|g_{kk}|^2 \\ &= \frac{\rho_f}{\mathrm{tr}(\mathbf{W}^{-1})}\end{aligned} \tag{6}$$

Based on (6), the sum rate achieved by the ZF precoding can be presented as [5]

$$C_{ZF} = \sum_{i=1}^{K}\log_2(1+\gamma_i) = K\log_2\left(1+\frac{\rho_f}{\mathrm{tr}\left(\mathbf{W}^{-1}\right)}\right). \tag{7}$$

For downlink large-scale MIMO systems, it has been verified that the ZF precoding can achieve the capacity close to

the optimal DPC precoding [5]. However, the ZF precoding involves matrix inversion $\mathbf{W}^{-1}$ of large size, and the computational complexity of $\mathbf{W}^{-1}$ is $\mathcal{O}(K^3)$, which is high since K is usually large in large-scale MIMO systems.

B. Linear precoding based on Gauss-Seidel method

Although the computation of $\mathbf{W}^{-1}$ is complicated, fortunately, the special channel properties of large-scale MIMO enable us to obtain the precoded signal vector $\mathbf{t}$ (or equivalently $\hat{\mathbf{s}}$) in (5) with low complexity. For downlink large-scale MIMO systems, the columns of channel matrix $\mathbf{H}$ are asymptotically orthogonal [5]. Therefore, we have $\mathbf{q}^H\mathbf{W}\mathbf{q} = \mathbf{q}^H\mathbf{H}(\mathbf{q}^H\mathbf{H})^H > 0$, where $\mathbf{q}$ is an arbitrary $N \times 1$ non-zero vector. This means matrix $\mathbf{W}$ is positive definite. Besides, since we have $\mathbf{W}^H = (\mathbf{H}\mathbf{H}^H)^H = \mathbf{W}$, we can conclude that matrix $\mathbf{W}$ is Hermitian positive definite.

The special property that $\mathbf{W}$ in (5) is a Hermitian positive definite matrix for downlink large-scale MIMO systems inspires us to exploit the GS method to efficiently solve (5) in an iterative way without matrix inversion. The GS method is used to solve the linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A}$ is the $N \times N$ Hermitian positive definite matrix, $\mathbf{x}$ is the $N \times 1$ solution vector, and $\mathbf{b}$ is the $N \times 1$ measurement vector. Unlike the traditional method that directly computes $\mathbf{A}^{-1}\mathbf{b}$ to obtain $\mathbf{x}$, the GS method can iteratively solve the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ with low complexity. Since matrix $\mathbf{A}$ is Hermitian positive definite, we can decompose $\mathbf{A}$ into a diagonal component $\mathbf{D}_\mathbf{A}$, a strictly lower triangular component $\mathbf{L}_\mathbf{A}$, and a strictly upper triangular component $\mathbf{L}_\mathbf{A}^H$. Then the solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ can be iteratively achieved by the GS method as [11]

$$\mathbf{x}^{(i+1)} = (\mathbf{D}_\mathbf{A} + \mathbf{L}_\mathbf{A})^{-1}(\mathbf{b} - \mathbf{L}_\mathbf{A}^H\mathbf{x}^{(i)}), \quad i = 0, 1, 2, \cdots \quad (8)$$

where the superscript i denotes the number of iterations.

Due to $\mathbf{W}$ is Hermitian positive definite as mentioned above, we can also decompose $\mathbf{W}$ as

$$\mathbf{W} = \mathbf{D} + \mathbf{L} + \mathbf{L}^H, \quad (9)$$

where $\mathbf{D}$, $\mathbf{L}$, and $\mathbf{L}^H$ denote the diagonal component, the strictly lower triangular component, and the strictly upper triangular component of $\mathbf{W}$, respectively. Then we can exploit the GS method to approximate $\hat{\mathbf{s}} = \mathbf{W}^{-1}\mathbf{s}$ in (5) as below

$$\hat{\mathbf{s}}^{(i+1)} = (\mathbf{D} + \mathbf{L})^{-1}(\mathbf{s} - \mathbf{L}^H\hat{\mathbf{s}}^{(i)}), \quad i = 0, 1, 2, \cdots \quad (10)$$

where $\hat{\mathbf{s}}^{(0)}$ denotes the initial solution, which is usually set as a $K \times 1$ zero vector without loss of generality [11]. Then the precoded signal vector for transmission can be achieved by

$$\mathbf{t} = \beta\mathbf{H}^H\hat{\mathbf{s}}^{(i+1)}. \quad (11)$$

As $(\mathbf{D} + \mathbf{L})$ is a lower triangular matrix, one can solve (10) to obtain $\hat{\mathbf{s}}^{(i+1)}$ with low complexity as will be addressed in Section III-D. It is worth noting that the proposed GS-based precoding scheme is convergent for any initial solution since the matrix $\mathbf{W}$ is Hermitian positive definite [11, Theorem 7.2.2]. Next we will prove that the GS-based precoding scheme has a faster convergence rate than the recently proposed Neumann-based precoding scheme [10].

C. Convergence rate

In this part, the approximation error induced by the GS method is analyzed at first. From (10), we can observe that the approximation error can be presented as

$$\hat{\mathbf{s}}^{(i+1)} - \hat{\mathbf{s}} = \mathbf{B}_G\left(\hat{\mathbf{s}}^{(i)} - \hat{\mathbf{s}}\right) = \cdots = \mathbf{B}_G^{i+1}\left(\hat{\mathbf{s}}^{(0)} - \hat{\mathbf{s}}\right). \quad (12)$$

where $\mathbf{B}_G = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{L}^H$ denotes the iteration matrix of the GS method. The approximation error can be evaluated as

$$\begin{aligned}\left\|\hat{\mathbf{s}}^{(i+1)} - \hat{\mathbf{s}}\right\|_2 &= \left\|\mathbf{B}_G^{i+1}\right\|_F\left\|\hat{\mathbf{s}}^{(0)} - \hat{\mathbf{s}}\right\|_2 \\ &\le \|\mathbf{B}_G\|_F^{(i+1)}\left\|\hat{\mathbf{s}}^{(0)} - \hat{\mathbf{s}}\right\|_2,\end{aligned} \quad (13)$$

which indicates that the final approximation error resulting from the GS method is affected by the Frobenius norm of the iteration matrix $\mathbf{B}_G$, and a smaller $\|\mathbf{B}_G\|_F$ will lead to a faster convergence rate [14]. The following Lemma 1 will verify that $\|\mathbf{B}_G\|_F$ is smaller than the Frobenius norm of the iteration matrix of the Neumann-based precoding scheme [10], which means the GS-based precoding scheme can enjoy a faster convergence rate.

Lemma 1. *For downlink large-scale MIMO systems, we have* $\|\mathbf{B}_G\|_F \le \frac{\|\mathbf{B}_N\|_F}{\sqrt{2}}$, *where* $\mathbf{B}_G = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{L}^H$ *and* $\mathbf{B}_N = \mathbf{D}^{-1}\left(\mathbf{L} + \mathbf{L}^H\right)$ *are the iteration matrices of the GS-based precoding and Neumann-based precoding schemes, respectively*.

Proof: Note that $\mathbf{B}_G$ can be rewritten as

$$\mathbf{B}_G = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{L}^H = -(\mathbf{I}_K + \mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{D}^{-1}\mathbf{L}^H. \quad (14)$$

According to the random matrix theory [5], for downlink large-scale MIMO systems with $N \gg K$, the elements of the diagonal matrix $\mathbf{D}$ will converge to N, while the elements of $\mathbf{L}$ will converge to 0. Thus, we have $\lim_{k\to\infty}\mathbf{D}^{-1}\mathbf{L} = \mathbf{0}$. Then the matrix $(\mathbf{I}_K + \mathbf{D}^{-1}\mathbf{L})^{-1}$ can be expanded as [14, Theorem 2.2.3]

$$\begin{aligned}(\mathbf{I}_K + \mathbf{D}^{-1}\mathbf{L})^{-1} &= \sum_{k=0}^{\infty}(-1)^k\left(\mathbf{D}^{-1}\mathbf{L}\right)^k \\ &= \mathbf{I}_k - \mathbf{D}^{-1}\mathbf{L} + \sum_{k=2}^{\infty}(-1)^k\left(\mathbf{D}^{-1}\mathbf{L}\right)^k.\end{aligned} \quad (15)$$

By considering the fact that $(\mathbf{D}^{-1}\mathbf{L})^k \to 0$ for a relatively large k (e.g., $k = 2$), we can only keep the first two items of the sum in (15) to approximate the matrix as $(\mathbf{I}_K + \mathbf{D}^{-1}\mathbf{L})^{-1} \approx \mathbf{I}_k - \mathbf{D}^{-1}\mathbf{L}$. Then, the iteration matrix $\mathbf{B}_G$ of the GS-based precoding scheme can be approached by

$$\mathbf{B}_G \approx \left(\mathbf{I}_k - \mathbf{D}^{-1}\mathbf{L}\right)\mathbf{D}^{-1}\mathbf{L}^H = \mathbf{D}^{-1}\mathbf{L}^H - \mathbf{D}^{-2}\mathbf{L}\mathbf{L}^H. \quad (16)$$

After this approximation, the Frobenius norm of $\mathbf{B}_G$ can be presented as

$$\begin{aligned}\|\mathbf{B}_G\|_F &\approx \left\|\mathbf{D}^{-1}\mathbf{L}^H - \mathbf{D}^{-2}\mathbf{L}\mathbf{L}^H\right\|_F \\ &\le \left\|\mathbf{D}^{-1}\mathbf{L}^H\right\|_F + \left\|\mathbf{D}^{-2}\mathbf{L}\mathbf{L}^H\right\|_F.\end{aligned} \quad (17)$$

Similar to the analysis above, the second item $\left\|\mathbf{D}^{-2}\mathbf{L}\mathbf{L}^H\right\|_F$ on the right hand of the inequality (17)

has a very limited contribution to $\|\mathbf{B}_G\|_F$, due to the fact that the elements of $\mathbf{D}$ are large while the elements of $\mathbf{L}$ are close to zero. Thus, we abandon this item and $\|\mathbf{B}_G\|_F$ can be upper bounded by $\left\|\mathbf{D}^{-1}\mathbf{L}^H\right\|_F$ as

$$\|\mathbf{B}_G\|_F \leq \left\|\mathbf{D}^{-1}\mathbf{L}^H\right\|_F = \left(\sum_{m=1}^{K}\sum_{k=1,k<m}^{K}\left|\frac{w_{mk}}{w_{mm}}\right|^2\right)^{\frac{1}{2}}. \quad (18)$$

Since the iteration matrix of the Neumann-based precoding scheme is $\mathbf{B}_N = \mathbf{D}^{-1}\left(\mathbf{L}+\mathbf{L}^H\right)$, the Frobenius norm of $\mathbf{B}_N$ can be obtained as

$$\begin{aligned}\|\mathbf{B}_N\|_F &= \left(\sum_{m=1}^{K}\sum_{k=1,k\neq m}^{K}\left|\frac{w_{mk}}{w_{mm}}\right|^2\right)^{1/2} \\ &= \left(2\sum_{m=1}^{K}\sum_{k=1,k<m}^{K}\left|\frac{w_{mk}}{w_{mm}}\right|^2\right)^{1/2} \\ &= \sqrt{2}\left\|\mathbf{D}^{-1}\mathbf{L}^H\right\|_F.\end{aligned} \quad (19)$$

According to (19), we can conclude that $\|\mathbf{B}_G\|_F \leq \frac{\|\mathbf{B}_N\|_F}{\sqrt{2}}$. ■

The **Lemma 1** indicates that the GS-based precoding scheme will enjoy a faster convergence rate than the Neumann-based precoding since it has a smaller $\|\mathbf{B}_G\|_F$ [14]. In other words, when the number of iterations is limited, the solution of the GS-based precoding will be more close to that of the ZF precoding with exact matrix inversion.

D. Computational complexity analysis

Since both the conventional ZF precoding and the proposed GS-based precoding need to compute $\mathbf{W}=\mathbf{H}\mathbf{H}^H$, we compare the computational complexity after the matrix $\mathbf{W}$ has been obtained. Besides, as the computational complexity is dominated by multiplications, we evaluate the complexity in terms of the required number of complex multiplications [15]. It can be found from (10) and (11) that the computational complexity of the proposed GS-based precoding scheme comes from three parts. The first one originates from solving the linear equation (10). Considering the definition of $\mathbf{D}$ and $\mathbf{L}$ in (9), the solution can be presented as

$$\begin{aligned}\hat{s}_m^{(i+1)} = \tfrac{1}{w_{mm}}(s_m - \sum_{k<m} w_{mk}\hat{s}_k^{(i+1)} - \sum_{k>m} w_{mk}\hat{s}_k^{(i)}), \\ m,k = 1,2,\cdots,K,\end{aligned} \quad (20)$$

where $\hat{s}_m$, $\hat{s}_m^{(i+1)}$, and s_m denote the mth element of $\hat{\mathbf{s}}^{(i)}$, $\hat{\mathbf{s}}^{(i+1)}$, and $\mathbf{s}$, respectively, w_{mk} denotes the entry of the matrix $\mathbf{W}$ in the mth row and kth column. It is clear that the required number of multiplications in the computation of $\hat{s}_m^{(i+1)}$ is K. Since there are K elements in $\hat{\mathbf{s}}^{(i+1)}$, solving the equation (10) only requires K^2 times of multiplications. The second one comes from the multiplication of a $N\times K$ matrix $\mathbf{H}^H$ and a $K\times 1$ vector $\hat{\mathbf{s}}^{(i+1)}$, where NK times of multiplications are required. The last one is from the computation of the factor β in (4). It has been proved that when N and K go infinity while $\alpha = N/K$ keeps fixed in large-scale MIMO systems, β converges to a deterministic value $\sqrt{K(\alpha-1)}$ [5]. Fig. 2 shows the comparison between the theoretical and simulated β against different α when K is fixed to 16. We can conclude from Fig. 2 that although N and K are finite in practical large-scale MIMO systems, the gap between the theoretical and simulated β is negligible except when $\alpha = 1$. Thus, once the configuration of the large-scale MIMO system has been fixed, the constant factor β is known, and we only need N times of multiplications to compute $\beta\mathbf{H}^H\hat{\mathbf{s}}^{(i+1)}$. Based on the analysis above, the overall required number of multiplications by the proposed GS-based precoding is $N+NK+iK^2$.

TABLE I
COMPUTATIONAL COMPLEXITY

Iterative number	Neumann-based precoding [10]	Proposed GS-based precoding
$i=2$	$N+NK+3K^2-K$	$N+NK+2K^2$
$i=3$	$N+NK+K^2+K^3$	$N+NK+3K^2$
$i=4$	$N+NK+2K^3$	$N+NK+4K^2$
$i=5$	$N+NK+3K^3-K^2$	$N+NK+5K^2$
$i=6$	$N+NK+4K^3-2K^2$	$N+NK+6K^2$
$i=7$	$N+NK+5K^3-3K^2$	$N+NK+7K^2$

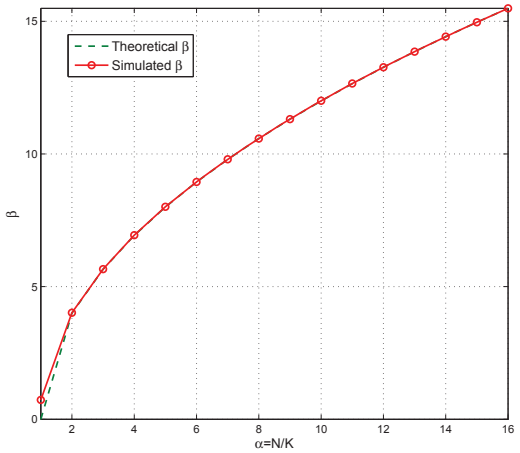

Fig. 2. Comparison between the theoretical and simulated β against different $\alpha = N/K$ for fixed $K = 16$ in large-scale MIMO systems.

Table I compares the complexity of the recently proposed Neumann-based precoding [10] and the proposed GS-based precoding. Since the complexity of the classical ZF precoding is $\mathcal{O}(K^3)$, we can conclude from Table I that the Neumann-based precoding can reduce the complexity from $\mathcal{O}(K^3)$ to $\mathcal{O}(K^2)$ when the number of iterations is $i=2$, but its complexity is still $\mathcal{O}(K^3)$ when $i\geq 3$. To ensure the approximation performance, usually a large number of iterations is required (as will be verified later in Section IV), which means the required number of multiplications by the Neumann-based precoding may be even larger than the ZF precoding. Therefore, although it does not require any division operation which is difficult for hardware implementation [5], [10], its overall complexity is almost the same as the ZF precoding. On the other hand, we can observe that the proposed GS-based precoding also requires no division operation since $\frac{1}{w_{mm}}$ can be approached by $\frac{1}{N}$ [5], and its complexity is $\mathcal{O}(K^2)$ for an

arbitrary number of iterations. Even for $i = 2$, the proposed GS-based precoding has lower complexity than the Neumann-based one [10].

Additionally, we can observe from (20) that the computation of $\hat{s}_m^{(i+1)}$ utilizes $\hat{s}_k^{(i+1)}$ for $k = 1, 2, \cdots, m-1$ in the current $(i+1)$th iteration and $s_l^{(i)}$ for $l = m+1, m+2, \cdots, K$ in the previous ith iteration. Then two another benefits can be expected. Firstly, after $\hat{s}_m^{(i+1)}$ has been obtained, we can use it to overwrite $\hat{s}_m^{(i)}$ which is useless in the next computation of $\hat{s}_{m+1}^{(i+1)}$. In this way, only one storage vector of size $K \times 1$ is required; Secondly, the solution to (10) becomes closer to the final solution $\hat{\mathbf{s}}$ with an increasing i, so $\hat{s}_m^{(i+1)}$ can exploit the elements $\hat{s}_k^{(i+1)}$ for $k = 1, 2, \cdots, m-1$ that have already been computed in current $(i+1)$th iteration to produce more reliable result than the Neumann-based precoding, which only utilizes all the elements of $\hat{\mathbf{s}}^{(i)}$ in the previous ith iteration. Thus, a faster convergence rate can be expected from another aspect, and the required number of iterations to achieve a certain approximation accuracy becomes smaller. Based on these two special advantages of the GS method, the overall complexity of the proposed scheme can be reduced further.

IV. Simulation Results

To evaluate the performance of the proposed GS-based precoding, we provide the simulation results of the achievable channel capacity as well as the BER performance, compared with the recently proposed Neumann-based precoding [10]. The capacity and BER performance of the classical MF precoding and ZF precoding with exact matrix inversion is also included as the benchmark for comparison. Besides, we also provide the performance of the optimal DPC precoding to verify the capacity-approaching performance of the proposed GS-based precoding. We consider two typical downlink large-scale MIMO configurations with $N \times K = 256 \times 16$ and $N \times K = 256 \times 32$, respectively. The modulation scheme of 64 QAM is employed, and the Rayleigh fading channel models is used for simulation.

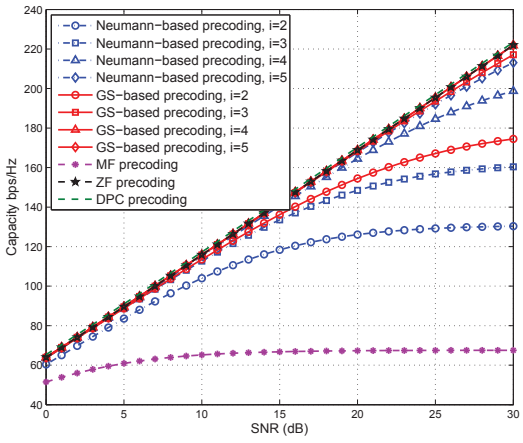

Fig. 3. Capacity comparison for the 256×16 downlink large-scale MIMO system.

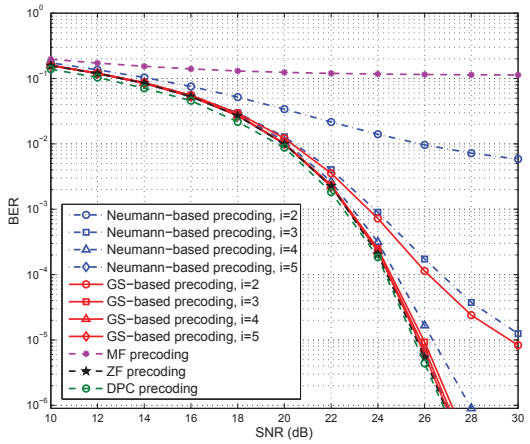

Fig. 4. BER performance comparison for the 256×16 downlink large-scale MIMO system.

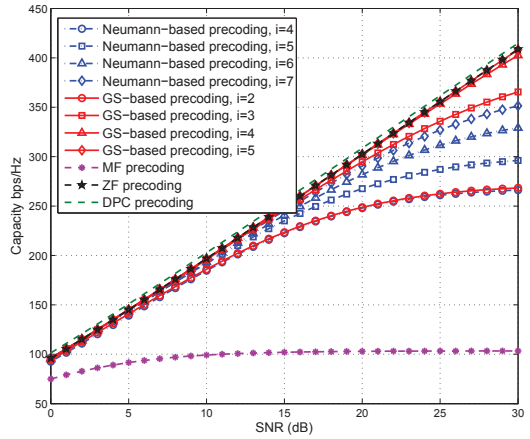

Fig. 5. Capacity comparison for the 256×32 downlink large-scale MIMO system.

Fig. 3 and Fig. 4 show the capacity and BER performance comparison between the Neumann-based precoding and GS-based precoding, respectively. The MIMO configuration is $N \times K = 256 \times 16$, and i denotes the number of iterations. It is clear from Fig. 3 that for a realistic large-scale MIMO system with limited number of transmit antennas and users, the ZF precoding has much better performance than the MF precoding, and it is capacity-approaching compared to the optimal DPC precoding, since the performance gap is within 0.5 dB for the achieved capacity of 220 bps/Hz. In addition, as shown in Fig. 3, when the number of iterations is small i.e., $i = 2$, the Neumann-based precoding cannot converge, leading to the serious multi-user interferences and the obvious loss in capacity, while the proposed GS-based precoding can achieve the much better performance. For example, when SNR = 30 dB, the proposed scheme can achieve 175 bps/Hz, while only 130 bps/Hz can be obtained by the Neumann-baesed precoding. As the number of iterations increases, the performance of both schemes improves. However, when the same number of iterations i is used, the proposed scheme

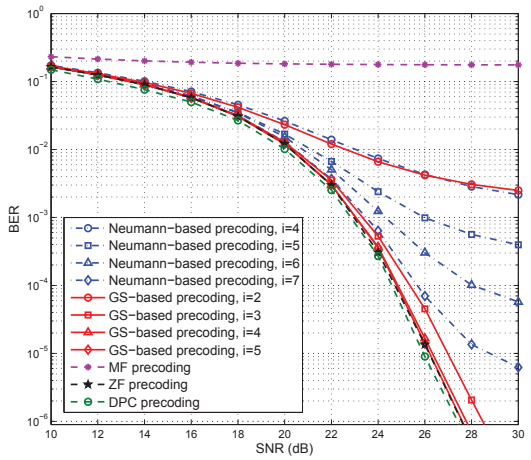

Fig. 6. BER performance comparison for the 256×32 downlink large-scale MIMO system.

outperforms the Neumann-based one. For example, we can observe from Fig. 4 that when $i = 3$, the required SNR to achieve the BER of 10^{-3} by the proposed scheme is 22.5 dB, while the Neumann-based one requires the SNR of 24 dB.

Fig. 5 and Fig. 6 show the capacity and BER performance comparison between the two precoding schemes when $N \times K = 256 \times 32$, respectively. Comparing Fig. 3 and Fig. 5, we can find that with a decreasing value of $\alpha = N/K$, the performance of Neumann-based precoding becomes worse. For example, when $i = 4$, for the 256×16 MIMO system, the Neumann-based precoding can achieve 90% of capacity of DPC precoding at SNR = 30 dB, while for the 256×32 MIMO system, it can only achieve 64% of the ideal capacity. In contrast, when $i = 4$, the proposed GS-based precoding can achieve 99% and 97% of capacity of DPC precoding for 256×16 and 256×32 MIMO systems, respectively. This indicates that the proposed scheme is more robust to α.

Moreover, we can also conclude from Fig. 6 that the proposed GS-based precoding requires a smaller number of iterations to obtain the same BER performance than the Neumann-based precoding. For example, when $i = 2$, the performance of the proposed scheme is almost the same as that of the Neumann-based one when $i = 4$, which means a faster convergence rate can be achieved by the proposed scheme as we have proved in Section III-C. Therefore, the complexity can be further reduced.

More importantly, when the number of iterations is relatively large (e.g., $i = 4$ in Fig. 3 or $i = 5$ in Fig. 5), the proposed scheme without the complicated matrix inversion can approach the channel capacity of the optimal DPC precoding with negligible performance loss, which verify the capacity-approaching performance of the proposed GS-based precoding.

V. Conclusions

In this paper, by fully exploiting the special channel property of downlink large-scale MIMO systems, we proposed a capacity-approaching GS-based precoding scheme with low complexity. We also prove that the proposed scheme can converge faster than the recently proposed Neumann-based precoding. It is shown that the proposed GS-based precoding can reduce the complexity from $\mathcal{O}(K^3)$ to $\mathcal{O}(K^2)$. Simulation results demonstrate that the proposed scheme outperforms the Neumann-based precoding, and approaches the optimal performance of the DPC precoding with a small number of iterations, i.e., $i = 4$ and $i = 5$ for $N \times K = 256 \times 16$ and $N \times K = 256 \times 32$ large-scale MIMO systems, respectively.

Acknowledgments

This work was supported by National Key Basic Research Program of China (Grant No. 2013CB329203), National High Technology Research and Development Program of China (Grant No. 2014AA01A704), National Nature Science Foundation of China (Grant Nos. 61271266 and 61201185), Science and Technology Foundation for Beijing Outstanding Doctoral Dissertation Supervisor (Grant No. 20121000303), and Foundation of Shenzhen government.

References

[1] R. W. Health, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[2] L. Dai, Z. Wang, and Z. Yang, "Next-generation digital television terrestrial broadcasting systems: Key technologies and research trends," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 150–158, Jun. 2012.

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[4] H. Ngo, E. Larsson, and T. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2012.

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

[6] M. H. Costa, "Writing on dirty paper (corresp.)," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, Mar. 1983.

[7] A. Razi, D. J. Ryan, I. B. Collings, and J. Yuan, "Sum rates, rate allocation, and user scheduling for multi-user MIMO vector perturbation precoding," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 356–365, Jan. 2010.

[8] J. H. Lee, "Lattice precoding and pre-distorted constellation in degraded broadcast channel with finite input alphabets," *IEEE Trans. Commun.*, vol. 58, no. 5, pp. 1315–1320, May 2010.

[9] Y. Wu, M. Wang, C. Xiao, Z. Ding, and X. Gao, "Linear precoding for MIMO broadcast channels with finite-alphabet constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2906–2920, Aug. 2012.

[10] H. Prabhu, J. Rodrigues, O. Edfors, and F. Rusek, "Approximative matrix inverse computations for large-scale MIMO and applications to linear pre-coding systems," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC'13)*, Apr. 2013, pp. 2710–2715.

[11] A. Björck, *Numerical Methods for Least Squares Problems*. Society for Industrial and Applied Mathematics (SIAM), 1996.

[12] 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. 1932–4553, May 2014.

[13] L. Dai, Z. Wang, and Z. Yang, "Spectrally efficient time-frequency training OFDM for mobile large-scale MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 251–263, Feb. 2013.

[14] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012.

[15] L. Dai, X. Gao, X. Su, S. Han, C.-L. I, and Z. Wang, "Low-complexity soft-output signal detection based on Gauss-Seidel method for uplink multi-user large-scale MIMO systems," *IEEE Trans. Veh. Technol.*, 2015.