Antenna Grouping based Feedback Reduction for FDD-based Massive MIMO Systems

Byungju Lee, Junil Choi, Ji-yun Seol, David J. Love, and Byonghyo Shim

***Abstract*—Recent works on massive multiple-input multiple-output (MIMO) have shown that a potential breakthrough in capacity gains can be achieved by deploying a very large number of antennas at the basestation. In order to achieve optimal performance of massive MIMO systems, accurate transmit-side channel state information (CSI) should be available at the basestation. While transmit-side CSI can be obtained by employing channel reciprocity in time division duplexing (TDD) systems, explicit feedback of CSI from the user terminal to the basestation is needed for frequency division duplexing (FDD) systems. In this paper, we propose an antenna grouping based feedback reduction technique for FDD-based massive MIMO systems. The proposed algorithm, dubbed antenna group beamforming (AGB), maps multiple correlated antenna elements to a single representative value using pre-designed patterns. The proposed method introduces the concept of using a *header* of overall feedback resources to select a suitable group pattern and the *payload* to quantize the reduced dimension channel vector. Simulation results show that the proposed method achieves significant feedback overhead reduction over conventional approach performing the vector quantization of whole channel vector under the same target sum rate requirement.**

***Index Terms*—Massive multiple-input multiple-output, antenna group beamforming, feedback reduction, vector quantization, Grassmannian subspace packing.**

I. Introduction

Multiple-input multiple-output (MIMO) systems with large-scale transmit antenna array, often called massive MIMO, have been of great interest in recent years because of their potential to dramatically improve spectral efficiency of future wireless systems [2], [3]. By employing tens or hundreds of antennas at the basestation, massive MIMO systems can control intra-cell interference and thermal noise by simply using linear precoding in the downlink and receive filtering in the uplink [2]. Additionally, massive MIMO can improve the power efficiency by scaling down the transmit power of each terminal inversely proportional to the number of basestation antennas [3].

Presently, standardization activity for massive MIMO has been initiated [4], [5], and there is on-going debate regarding the pros and cons of time division duplexing (TDD) and frequency division duplexing (FDD). In obtaining the CSI, FDD requires the CSI to be fed back through the uplink [6] while no such procedure is required for TDD systems owing to channel reciprocity [7]. In fact, under the assumption that RF chains are properly calibrated [8], the CSI of the downlink can be estimated using the pilot signal in the uplink so that the CSI feedback information from the user terminal to the basestation is unnecessary. Due to this benefit, most of the massive MIMO works in the literature have focused on TDD [9] (possible exceptions are [10], [11]). However, since FDD dominates current cellular networks and offers many benefits over TDD (e.g., small latency, continuous channel estimation, backward compatibility), it is important to identify and develop solutions for potential issues arising from FDD-based massive MIMO techniques.

One well-known problem of FDD system is that the amount of CSI feedback must scale linearly with the number of antennas to control the quantization error [12]–[14]. Therefore, it is not hard to convince oneself that the overhead of CSI feedback is a serious concern when the number of transmit antennas is large. Needless to say, a technique that efficiently reduces the feedback overhead while affecting minimal impact on system performance is crucial to the success of massive MIMO systems.

In this paper, we provide a novel framework for FDD-based massive MIMO systems that achieves a reduction in the CSI feedback overhead by exploiting the spatial correlation among antennas. The proposed algorithm, henceforth dubbed antenna group beamforming (AGB), maps multiple correlated antenna elements to a single representative value using properly designed grouping patterns. When the antenna elements are correlated, the loss caused by the grouping antenna elements is shown to be small, meaning that grouping of antenna elements with correlated channels is an effective means to reduce the dimension of the channel vector to be quantized. In fact, by allocating a small portion of the feedback resources to represent the grouping pattern, the number of bits required for channel vector quantization can be reduced substantially, thereby achieving significant reduction in feedback overhead. In order to support the antenna grouping operation, the proposed AGB algorithm divides the feedback resources into two parts: a *header* to indicate

B. Lee and B. Shim are with Dept. of Electrical and Computer Engineering, Seoul National University, Seoul, Korea, J. Choi and D. J. Love are with School of Electrical and Computer Engineering, Purdue Univ., West Lafayette, IN, USA, and J. Seol is with Samsung Electronics Co., Ltd., Suwon, Korea.

This paper was presented in part at the International Conference on Communications (ICC), 2014 [1].

This work was sponsored by Communications Research Team (CRT), DMC R&D Center, Samsung Electronics Co. Ltd, the MSIP (Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013 (KCA-12-911-01-110) and the NRF grant funded by the Korea government (MEST) (No. 2012R1A2A2A01047510).

the antenna group pattern and a *payload* to indicate the codebook index of the reduced dimension channel vector. At the user terminal, the pair of group pattern and codeword minimizing the quantization distortion is chosen. Using the information delivered from the user terminal, the basestation reconstructs the full-dimensional channel vector and then performs transmit beamforming.

In our analysis, we show that when the transmit antenna elements are correlated the proposed AGB algorithm exhibits smaller quantization distortion than the conventional vector quantization employing a channel statistic-based codebook. This in turn implies that the number of quantization bits required to meet the certain level of the performance for the AGB algorithm is smaller than that of conventional vector quantization under the same level of quantization distortion. We confirm by simulation on realistic massive MIMO channels that the proposed AGB algorithm achieves up to 20%∼70% savings in feedback information over the conventional vector quantization under the same target sum rate requirement.

The remainder of this paper is organized as follows. In Section II, we briefly review the system model and the conventional beamforming technique. In Section III, we provide a detailed description of the proposed AGB algorithm and subspace packing based grouping pattern generation scheme. We present the simulation results in Section IV and present conclusions in Section V.

Notations: Lower and upper boldface symbols are used to denote vectors and matrices, respectively. The superscripts $(\cdot)^H$ and $(\cdot)^T$ denote Hermitian transpose and transpose, respectively. $\|\mathbf{X}\|$ and $\|\mathbf{X}\|_F$ are used as the two-norm and the Frobenius norm of a matrix $\mathbf{X}$, respectively. $E[\cdot]$ denotes the expectation operation, and $\mathcal{CN}(m,\sigma^2)$ indicates a complex Gaussian distribution with mean m and variance σ^2. $\mathrm{tr}(\cdot)$ is the trace operation, and $\mathrm{vec}(\mathbf{X})$ is the vectorization of matrix $\mathbf{X}$. Let $\mathbf{X}_\Lambda \in \mathbb{C}^{|\Lambda|\times|\Lambda|}$ denote a submatrix of $\mathbf{X}$ whose (i,j)-th entry is $\mathbf{X}(\Lambda(i),\Lambda(j))$ for $i,j=1,\ldots,|\Lambda|$ (Λ is the set of partial indices and $|\Lambda|$ is the cardinality of Λ).

II. MIMO Beamforming

A. System Model and Conventional Beamforming

We consider a multiple-input single-output (MISO) downlink channel with N_t antennas at the basestation and K user terminals each with a single antenna (see Fig. 1). In this setup, the received signal at the k-th user can be expressed as

$$y_k[n] = \mathbf{h}_k^H \mathbf{w}_k s_k[n] + \mathbf{h}_k^H \sum_{j\neq k} \mathbf{w}_j s_j[n] + z_k[n] \qquad (1)$$

where $\mathbf{h}_k \in \mathbb{C}^{N_t}$ is the channel vector from the basestation antenna array to the k-th user, $\mathbf{w}_i \in \mathbb{C}^{N_t}$ is the unit norm beamforming vector ($\|\mathbf{w}_i\|^2 = 1$), and $s_i[n] \in \mathbb{C}$ is the message signal for the i-th user. When the channels are spatially correlated across the transmit antenna array, the channel vector can be readily modeled as $\mathbf{h}_k = \mathbf{R}_{t,k}^{1/2}\mathbf{h}_{w,k}$

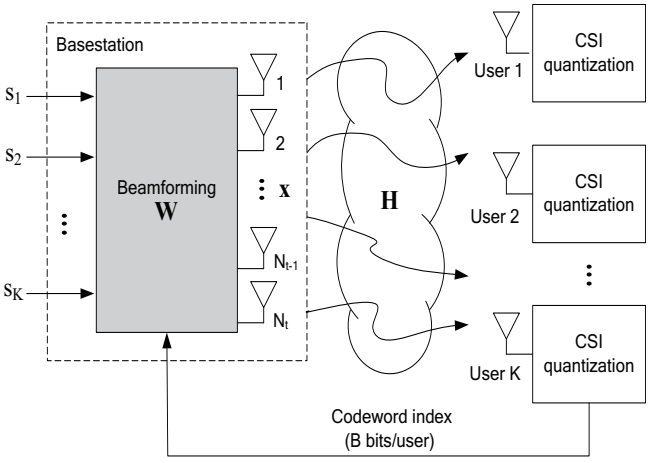


Fig. 1. CSI feedback in the multi-user downlink system.

where $\mathbf{R}_{t,k} \in \mathbb{C}^{N_t\times N_t}$ is the transmit correlation matrix of the k-th user and $\mathbf{h}_{w,k} \in \mathbb{C}^{N_t}$ is the channel vector whose elements are independent and identically distributed according to $\mathbf{h}_{w,k} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_t})$ [15]. The matrix-vector form of (1) is expressed as

$$\mathbf{y}[n] = \mathbf{H}\mathbf{x}[n] + \mathbf{z}[n] \qquad (2)$$

where $\mathbf{H} = [\mathbf{h}_1\ \mathbf{h}_2\ \ldots\ \mathbf{h}_K]^H \in \mathbb{C}^{K\times N_t}$ is the composite channel matrix, $\mathbf{z}[n] = [z_1[n]\ \ z_2[n]\ \ldots\ \ z_K[n]]^T \in \mathbb{C}^K$ is the complex Gaussian noise vector ($\mathbf{z}[n] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_K)$), and $\mathbf{x}[n]$ and $\mathbf{y}[n] = [y_1[n]\ y_2[n]\ \ldots,\ y_K[n]]^T$ are the transmit vector normalized with the power constraint ($E[\|\mathbf{x}[n]\|^2] = P$) and the vectorized received signal vector, respectively. In order to control the inter-user interference, beamforming is applied and hence $\mathbf{x}[n] = \mathbf{W}\mathbf{s}[n]$ where $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2\ \ldots\ \mathbf{w}_K] \in \mathbb{C}^{N_t\times K}$ and $\mathbf{s}[n] = [s_1[n]\ s_2[n]\ \ldots\ s_K[n]]^T \in \mathbb{C}^K$ are the beamforming matrix and the message vector, respectively. Under the assumption that the basestation allocates equal power for all users, the achievable rate of the k-th user is

$$R_k = \log_2\left(1 + \frac{\frac{P}{K}|\mathbf{h}_k^H\mathbf{w}_k|^2}{1 + \frac{P}{K}\sum_{j=1,j\neq k}^{K}|\mathbf{h}_k^H\mathbf{w}_j|^2}\right) \qquad (3)$$

and the corresponding sum rate becomes $R_{\mathrm{sum}} = \sum_{k=1}^{K} R_k$.

In generating the beamforming vectors, we consider zero-forcing beamforming (ZFBF) [16]–[18] where the right pseudo inverse $\mathbf{W}_{\mathrm{zf}} = \mathbf{H}^H\left(\mathbf{H}\mathbf{H}^H\right)^{-1}$ of the channel matrix $\mathbf{H}$ is applied to the message vector $\mathbf{s}[n]$ to remove the inter-user interference.[1] In order to satisfy the transmit power constraint, the beamforming vector $\mathbf{w}_k$ should be normalized as

$$\mathbf{w}_k = \frac{\mathbf{W}_{\mathrm{zf}}^k}{\|\mathbf{W}_{\mathrm{zf}}^k\|} \qquad (4)$$

[1]Note that one might consider the maximal ratio transmission (MRT) based beamforming (i.e., $\mathbf{w}_{\mathrm{mrt}} = \mathbf{H}^H$). Since the MRT precoding is in principle based on the law of large numbers, the inter-user interference control of the MRT beamforming will be effective only when excessive numbers of transmit antennas is used and further there is no correlation among antenna elements.

where $\mathbf{W}_{\text{zf}}^k$ is the k-th column of $\mathbf{W}_{\text{zf}}$. When perfect CSI is available, $\mathbf{h}_k^H \mathbf{w}_j = 0$ $(j \neq k)$ and hence the received signal at the k-th user becomes

$$y_k[n] = \mathbf{h}_k^H \mathbf{w}_k s_k[n] + z_k[n]. \tag{5}$$

The sum rate in (3) is simplified to $R_{\text{sum}} = \sum_{k=1}^{K} \log_2(1 + \frac{P}{K}|\mathbf{h}_k^H \mathbf{w}_k|^2)$.

In practical systems, however, the basestation uses the quantized channel matrix $\hat{\mathbf{H}}$ delivered from mobile terminals so that the composite channel matrix is $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \cdots, \hat{\mathbf{h}}_K]^H$ and the corresponding beamforming vector is $\hat{\mathbf{w}}_k = \hat{\mathbf{W}}_{\text{zf}}^k / \|\hat{\mathbf{W}}_{\text{zf}}^k\|$ where $\hat{\mathbf{W}}_{\text{zf}} = \hat{\mathbf{H}}^H \left(\hat{\mathbf{H}}\hat{\mathbf{H}}^H\right)^{-1}$. Due to the imperfect channel information, the inter-user interference $(\sum_{j=1, j\neq k}^{K} \mathbf{h}_k^H \hat{\mathbf{w}}_j)$ cannot be perfectly removed and hence the sum rate becomes

$$R_{\text{sum}} = \sum_{k=1}^{K} \log_2 \left(1 + \frac{\frac{P}{K}|\mathbf{h}_k^H \hat{\mathbf{w}}_k|^2}{1 + \frac{P}{K}\sum_{j=1, j\neq k}^{K} |\mathbf{h}_k^H \hat{\mathbf{w}}_j|^2}\right). \tag{6}$$

B. Limited Feedback

In order to feed back the CSI, a user quantizes its channel direction $\bar{\mathbf{h}}_k = \frac{\mathbf{h}_k}{\|\mathbf{h}_k\|}$ to a unit norm vector $\hat{\mathbf{h}}_k$. Specifically, the k-th user chooses the quantized vector (codeword) $\hat{\mathbf{h}}_k$ from a pre-defined B-bit codebook set $\mathcal{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_{2^B}\}$ that is closest to its channel direction. That is,

$$\hat{\mathbf{h}}_k = \arg\max_{\mathbf{c}\in\mathcal{C}} |\bar{\mathbf{h}}_k^H \mathbf{c}|^2. \tag{7}$$

Then, the index of the chosen codeword $\hat{\mathbf{h}}_k$ is fed back to the basestation.

In general, the number of bits needed to express the codeword should be scaled with the dimension of the channel vector to be quantized to control the distortion caused by the quantization process. In particular, when there is no spatial correlation among antenna elements and the random vector quantization (RVQ) is used, the number of feedback bits per user should be scaled with the number of transmit antennas and SNR (in decibels) as [13], [19]

$$B_{\text{user}} = (N_t - 1)\log_2 P \approx \frac{N_t - 1}{3} P_{\text{dB}} \tag{8}$$

to maintain a constant gap in terms of the sum rate from the system with perfect CSI. Clearly, when the number of transmit antennas increases, the feedback overhead needs to be increased as well (e.g., $B_{\text{user}} = 210$ when $N_t = 64$, $P_{\text{dB}} = 10$), let alone the computational burden caused by the codebook selection. Apparently, the reduction of the channel vector dimension would be beneficial in reducing the feedback overhead of the FDD-based massive MIMO systems.

III. Antenna Grouping based Feedback Reduction

The key feature of the proposed scheme is to map multiple correlated antenna elements into a single representative value using pre-designed grouping patterns. As a result, the dimension of the channel vector is reduced and a codeword is chosen from the codebook generated by the reduced dimension channel vector. When the channel is correlated (i.e., antenna elements in a group are similar), the loss caused by the grouping antenna elements is negligible and the target performance can be achieved with smaller number of feedback bits than a conventional scheme requires. In this section, we explain the overall procedure of the proposed AGB algorithm and then discuss the pattern set design problem. We also analyze the quantization distortion caused by the AGB technique and show that the quantization distortion indeed decreases with the transmit correlation coefficient.

A. AGB Algorithm

As mentioned, the AGB algorithm achieves the reduction of the dimension in the channel vector to be quantized from N_t to N_g $(N_t > N_g)$ by mapping multiple correlated antenna elements to a single representative value (see Fig. 2). While a conventional scheme employs all feedback resources (B bits) to express the quantized channel vector, the proposed method uses part of the feedback resources to quantize the (reduced dimension) channel vector and the rest to express the grouping pattern. To support this, we divide the feedback resources into two parts: a header (B_p bits) to indicate the antenna group pattern and a payload (B-B_p bits) to represent an index of the quantized channel vector (see Fig. 3). Since the antenna grouping based quantization is performed separately per user, in the sequel we focus on the operation of a user and drop the user index k.

Suppose there are $N_P = 2^{B_p}$ antenna group patterns, then N_P distinct reduced dimension channel vectors are generated. Each pattern converts an N_t-dimensional channel vector into an N_g-dimensional vector by multiplying a grouping matrix $\mathbf{G}^{(i)} \in \mathbb{R}^{N_g \times N_t}$ to the channel vector. The reduced dimension channel vector $\mathbf{h}_r^{(i)} \in \mathbb{C}^{N_g}$ of the group pattern i is

$$\mathbf{h}_r^{(i)} = \mathbf{G}^{(i)}\mathbf{h}, \quad i = 1, \cdots, N_P. \tag{9}$$

In Fig. 4, we illustrate the concept of antenna group patterns for $N_t = 16$. One simple way to generate the reduced dimension channel vector is to average the channel coefficients in an antenna group. For example, if $\mathbf{h} = [\text{h}_1\ \text{h}_2\ \text{h}_3\ \text{h}_4]^T$ and two adjacent channel coefficients (first and second, third and fourth) are grouped, the mapping matrix is

$$\mathbf{G}^{(i)} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{10}$$

and $\mathbf{h}_r^{(i)} = \mathbf{G}^{(i)}\mathbf{h} = [\frac{\text{h}_1+\text{h}_2}{2} \quad \frac{\text{h}_3+\text{h}_4}{2}]^T$. One can alternatively consider a linear combination of antenna elements in

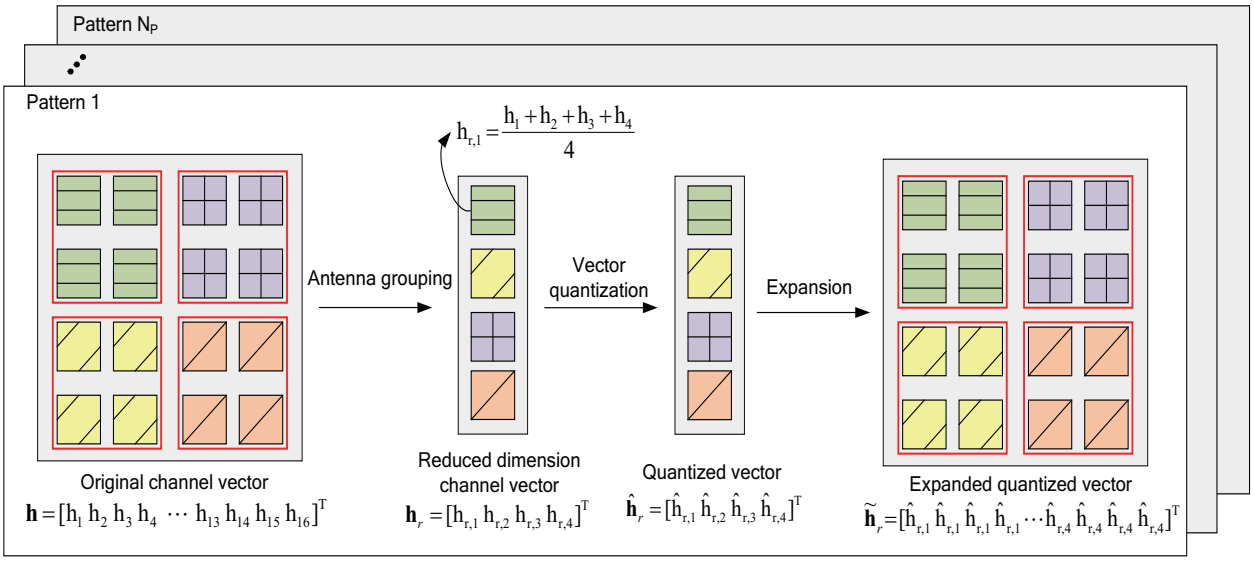


Fig. 2. Illustration of the AGB algorithm for $N_t = 8, N_g = 4$. The reduced dimension channel vector $\mathbf{h}_r$ is obtained by mapping antenna elements of a group as a representative value. Note that $\hat{\mathbf{h}}_r$ is the quantized version of $\mathbf{h}_r$ and $\tilde{\mathbf{h}}_r^{(i)}$ is expanded version of $\hat{\mathbf{h}}_r$.

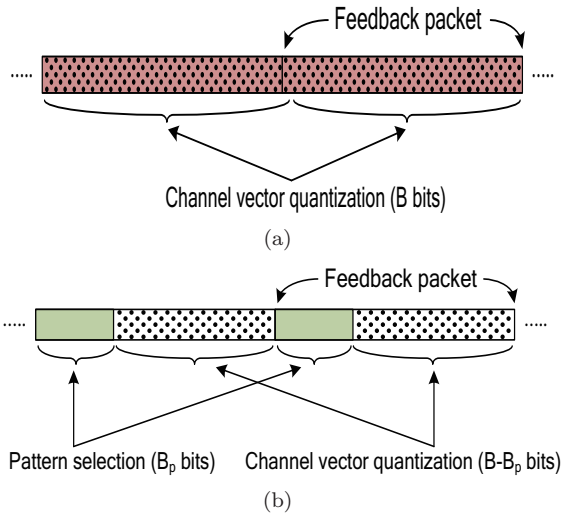


Fig. 3. Feedback packet structure: (a) conventional method and (b) proposed method.

each group. This approach, although it may offer better quantization error, will increase the feedback overhead due to the additional weight information.

Once the reduced dimension channel vector $\mathbf{h}_r^{(i)}$ is obtained, $\mathbf{h}_r^{(i)}$ is quantized by a $B - B_p$ bit codebook $\mathcal{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_{2^{B-B_P}}\}$. We note that since a codebook designed for i.i.d channels is not a proper choice for correlated channels, we consider a channel statistic-based codebook for channel vector quantization [20] (see Section III.C for details). The codeword $\hat{\mathbf{h}}_r^{(i)}$ maximizing the absolute inner product with $\mathbf{h}_r^{(i)}$ is chosen as

$$\hat{\mathbf{h}}_r^{(i)} = \arg\max_{\mathbf{c}\in\mathcal{C}} |\bar{\mathbf{h}}_r^{(i)H}\mathbf{c}|^2, \quad i = 1, \cdots, N_P \tag{11}$$

where $\bar{\mathbf{h}}_r^{(i)} = \frac{\mathbf{h}_r^{(i)}}{\|\mathbf{h}_r^{(i)}\|}$ is the direction of the reduced dimension channel vector for the i-th pattern. This process is repeated for each group pattern and N_p candidate codewords $\hat{\mathbf{h}}_r^{(i)}$, $i = 1, \cdots, N_P$, are chosen in total.

Once N_P candidate codewords are obtained, the codeword minimizing the distortion between $\bar{\mathbf{h}}$ and $\hat{\mathbf{h}}_r^{(i)}$ is chosen. We note that the comparison between $\hat{\mathbf{h}}$ and $\hat{\mathbf{h}}_r^{(i)}$ is not possible since the dimension of $\hat{\mathbf{h}}_r^{(i)} \in \mathbb{C}^{N_g}$ is smaller than that of the original channel vector $\mathbf{h} \in \mathbb{C}^{N_t}$. Thus, we use $\tilde{\mathbf{h}}_r^{(i)} \in \mathbb{C}^{N_t}$, an expanded version of $\hat{\mathbf{h}}_r^{(i)}$, to compute the distortion ($D(\mathbf{h}, \tilde{\mathbf{h}}_r^{(i)}) = E[\|\mathbf{h}\|^2(1-|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r^{(i)}|^2)]$) caused by the grouping and quantization. The expansion process, which essentially is done by copying each element in $\hat{\mathbf{h}}_r^{(i)}$ to $\frac{N_t}{N_g}$ elements in $\tilde{\mathbf{h}}_r^{(i)}$, is performed by multiplying an expansion matrix $\mathbf{E}^{(i)} \in \mathbb{R}^{N_t \times N_g}$ to $\hat{\mathbf{h}}_r^{(i)}$. The expanded quantized vector $\tilde{\mathbf{h}}_r^{(i)}$ is expressed as

$$\tilde{\mathbf{h}}_r^{(i)} = \mathbf{E}^{(i)}\hat{\mathbf{h}}_r^{(i)}, \quad i = 1, \cdots, N_P \tag{12}$$

where $\mathbf{E}^{(i)} = \kappa \left(\mathbf{G}^{(i)}\right)^T$ and satisfies $\mathbf{G}^{(i)}\mathbf{E}^{(i)} = \mathbf{I}_{N_g}$ ($\kappa = \frac{N_t}{N_g}$). For example, for the grouping matrix in (10), $\mathbf{E}^{(i)} = \kappa \left(\mathbf{G}^{(i)}\right)^T = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}^T$ and the expanded quantized vector is

$$\tilde{\mathbf{h}}_r^{(i)} = \mathbf{E}^{(i)}\hat{\mathbf{h}}_r^{(i)} = \kappa(\mathbf{G}^{(i)})^T\hat{\mathbf{h}}_r^{(i)} = \left[\hat{\mathrm{h}}_{r,1}^{(i)}\ \hat{\mathrm{h}}_{r,1}^{(i)}\ \hat{\mathrm{h}}_{r,2}^{(i)}\ \hat{\mathrm{h}}_{r,2}^{(i)}\right]^T. \tag{13}$$

The group pattern index i^* minimizing the distortion between $\mathbf{h}$ and $\tilde{\mathbf{h}}_r^{(i)}$ is

$$i^* = \arg\min_{i=1,\cdots,N_P} D(\mathbf{h}, \tilde{\mathbf{h}}_r^{(i)}). \tag{14}$$

Once this pattern index i^* is obtained, we deliver this index and the corresponding codeword index to the basestation. After receiving the pattern index and codeword index of all user terminals, the basestation decompresses the reduced dimension channel vector via the expansion ($\hat{\mathbf{h}} = \mathbf{E}^{(i^*)}\hat{\mathbf{h}}_r^{(i^*)}$) and then performs the beamforming using

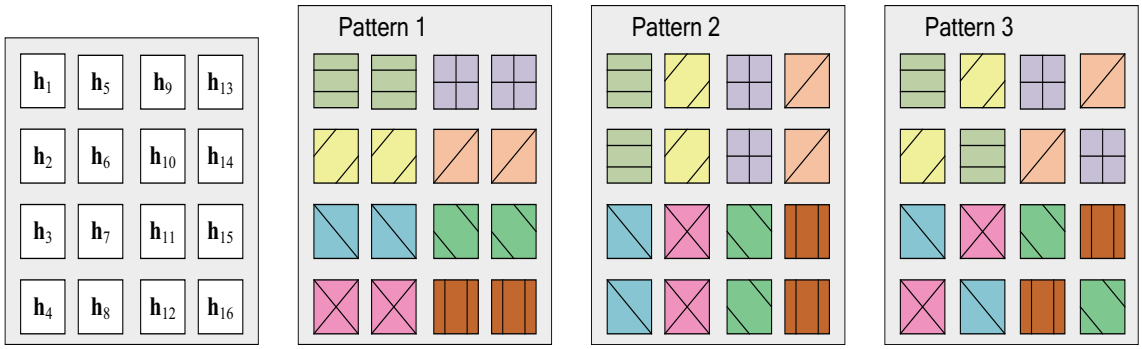


Fig. 4. Example of antenna group patterns ($N_t = 16, N_g = 8, N_P = 3$). Antenna elements belonging to the same pattern are mapped to one representative value.

the composite channel matrix $\hat{\mathbf{H}}$. A block diagram of the proposed AGB algorithm is depicted in Fig. 5.

B. Antenna Group Pattern Generation

Since multiple correlated antenna elements are mapped to a single representative value, the AGB algorithm is sensitive to the choice of the antenna group pattern. Without doubt, selecting the best pattern among all possible combinations would be the ideal option. However, since the number of patterns increases exponentially with the number of transmit antennas, it is not possible to investigate all possible patterns for the massive MIMO systems. This observation tells us that a simple yet effective pattern design is crucial to the success of the AGB algorithm.

One easy and intuitive way to construct an antenna group pattern $\mathbf{G}^{(i)}$ is to group highly correlated antenna elements together. Typically, adjacent antenna elements are highly correlated so that the grouping of nearby antenna elements may be a desirable option in practice (see the example in Fig. 4). Alternatively, one can consider Grassmannian subspace packing in the design of the antenna group patterns [21]. The main goal of Grassmannian subspace packing is, when the subspace distance metric and the number of feedback bits B are provided, to find a set of 2^B subspaces in[2] $\mathcal{G}(N_t, m)$ that maximizes the minimum subspace distance between any pair of subspaces in the set. The chordal distance metric has been popularly used in generating the beamforming codebook [21]. Our task of generating the pattern set is similar in spirit to the Grassmannian subspace packing based codebook generation in the sense that we construct a pattern set (containing 2^{B_p} patterns) from all possible pattern candidates ($\mathbf{E}^{(i)} \in \mathbb{R}^{N_t \times N_g}$) using a distance metric exploiting the spatial correlation among antenna elements.

In the first step of the pattern set design, we compute the quasi-correlation matrix norm $\|\tilde{\mathbf{R}}_t^{(i)}\|_F$ to measure the spatial proximity of the antenna elements in the antenna group. The quasi-correlation matrix $\tilde{\mathbf{R}}_t^{(i)}$, defined as $\tilde{\mathbf{R}}_t^{(i)} = \mathbf{R}_t^{1/2}\mathbf{E}^{(i)}$, captures the actual influence of the pattern $\mathbf{E}^{(i)}$ on the transmit correlation matrix $\mathbf{R}_t$. In general, a pattern generated by grouping closely spaced antenna elements tends to have a higher quasi-correlation matrix norm than that generated by grouping antenna elements apart. Thus, one can expect that a pattern with a large-correlation matrix norm exhibits lower grouping loss than that with a small quasi-correlation matrix norm. For patterns with high quasi-correlation matrix norm, we perform subspace packing to generate 2^{B_p} patterns (expansion matrices) maximizing the minimum distance metric between any pair of subspaces. To measure the distance, we use the correlation matrix distance $d_{\text{corr}}(\mathbf{A}, \mathbf{B})$ between two matrices $\mathbf{A}$ and $\mathbf{B}$ [22]

$$d_{\text{corr}}(\mathbf{A}, \mathbf{B}) = 1 - \frac{\text{tr}(\mathbf{A}^H\mathbf{B})}{\|\mathbf{A}\|_{\text{F}}\|\mathbf{B}\|_{\text{F}}}. \tag{15}$$

Note that $d_{\text{corr}}(\mathbf{A}, \mathbf{B})$ measures the orthogonality between two correlation matrices $\mathbf{A}$ and $\mathbf{B}$. When the correlation matrices are equal up to a scaling factor, d_{corr} is minimized ($d_{\text{corr}} = 0$). On the other hand, when the inner product between the vectorized correlation matrices is zero (i.e., vec($\mathbf{A}$) and vec($\mathbf{B}$) are orthogonal), d_{corr} is maximized ($d_{\text{corr}} = 1$). In our numerical simulations, we show that the proposed subspace packing approach achieves a substantial gain over an approach using randomly selected patterns (see Section IV.B). We summarize the antenna group pattern generation procedures in Table I.

C. Quantization Distortion Analysis

We now turn to the performance analysis of the AGB algorithm. In our analysis, we analyze the distortion D induced by the quantization of the channel direction vector $\bar{\mathbf{h}} = \frac{\mathbf{h}}{\|\mathbf{h}\|}$, which is defined as

$$\begin{aligned} D &= E\left[\|\mathbf{h}\|^2 - |\mathbf{h}^H\tilde{\mathbf{h}}_r|^2\right] \\ &= E\left[\|\mathbf{h}\|^2\left(1 - |\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right)\right] \end{aligned} \tag{16}$$

where $\tilde{\mathbf{h}}_r$ is the expanded version of the quantized vector $\hat{\mathbf{h}}_r$ (see (12)).

In order to evaluate the distortion D, we use the quantization cell upper bound (QUB) [23]. As mentioned, since a codebook designed for the i.i.d channels is not the right choice for correlated channels, we employ a channel statistic-based codebook obtained by applying

[2] $\mathcal{G}(N_t, m)$ is the set of m-dimensional subspaces in $\mathbb{C}^{N_t}$ (or $\mathbb{R}^{N_t}$).

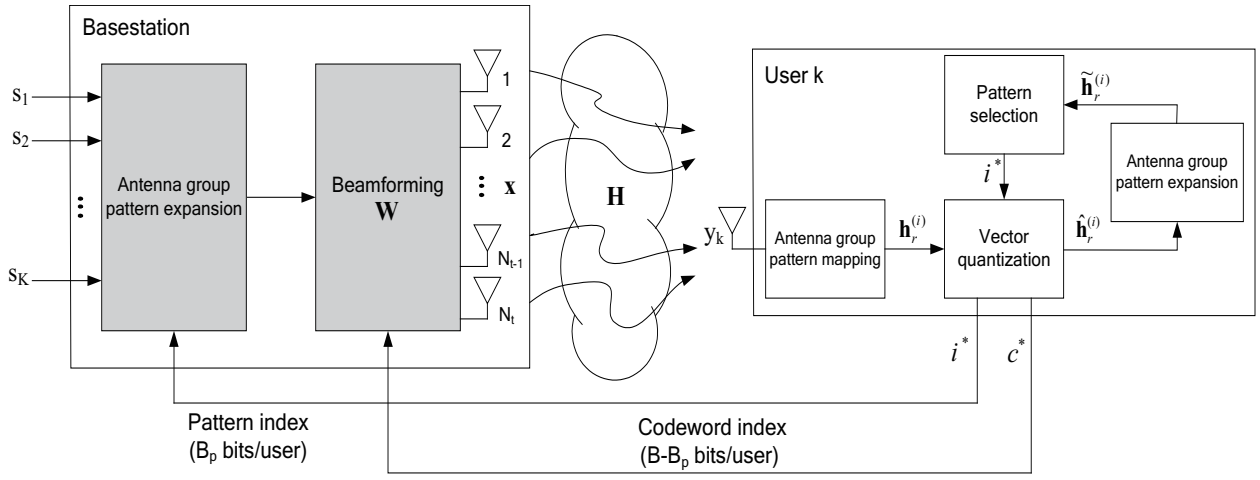


Fig. 5. A block diagram of the AGB algorithm in the multi-user downlink system.

the transmit correlation matrix $\mathbf{R}_t^{1/2}$ to the codebook generated from the GLP. Let $\mathbf{f}_i \in \mathbb{C}^r$ be the i-th unit norm vector generated from the GLP, then the set of B-bit codewords for the channel statistic-based codebook is [20]

$$\mathcal{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_{2^B}\} = \left\{ \frac{\mathbf{R}_t^{1/2}\mathbf{f}_1}{\|\mathbf{R}_t^{1/2}\mathbf{f}_1\|}, \cdots, \frac{\mathbf{R}_t^{1/2}\mathbf{f}_{2^B}}{\|\mathbf{R}_t^{1/2}\mathbf{f}_{2^B}\|} \right\}. \tag{17}$$

When the channel statistic-based codebook is used, the normalized distortion $\frac{D}{E[\|\mathbf{h}\|^2]} = 1 - |\bar{\mathbf{h}}^H\mathbf{c}_i|^2$ between the channel direction vector $\mathbf{h}$ and codebook vector $\mathbf{c}_i$ can be upper bounded as [15]

$$\mathcal{R}_i \approx \{\bar{\mathbf{h}} : 1 - |\bar{\mathbf{h}}^H\mathbf{c}_i|^2 \le \delta\} \tag{18}$$

where $\delta = \frac{\sigma_2^2}{\sigma_1^2} 2^{-\frac{B}{r-1}}$ (σ_i is the i-th largest singular value of the transmit correlation matrix $\mathbf{R}_t \in \mathbb{C}^{r\times r}$) and B is the number of quantization bits.

In our analysis, we restrict our attention to the scenario where two antenna elements are mapped to a single representative value for mathematical tractability. Nevertheless, since the key factor affecting the quantization distortion is the transmit correlation coefficient (see (29)), our results can be readily applied to the general scenario where more than two antenna elements are grouped together. The minimal set of assumptions used for the analytical tractability are as follows:

A-i) The channel vector $\bar{\mathbf{h}}$ is partitioned into two subvectors $\bar{\mathbf{h}}_\mathrm{A}$ and $\bar{\mathbf{h}}_\mathrm{B}$ ($\bar{\mathbf{h}} = [\bar{\mathbf{h}}_\mathrm{A}^T\ \bar{\mathbf{h}}_\mathrm{B}^T]^T$). $\bar{\mathbf{h}}_\mathrm{A}$ is a vector composed of N_g antenna elements and each entry of $\bar{\mathbf{h}}_\mathrm{A}$ is the first element in the group. Therefore, there are N_g antenna groups and each group consists of two antenna elements.

A-ii) The reduced dimension channel vector $\mathbf{h}_r$ is designed such that $\mathbf{h}_r = \bar{\mathbf{h}}_\mathrm{A}$. For example, if $\bar{\mathbf{h}} = [\bar{\mathrm{h}}_1\ \bar{\mathrm{h}}_2\ \bar{\mathrm{h}}_3\ \bar{\mathrm{h}}_4]^T$, then

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{19}$$

and hence $\mathbf{h}_r = \bar{\mathbf{h}}_\mathrm{A} = [\bar{\mathrm{h}}_1\ \bar{\mathrm{h}}_3]^T$ ($\bar{\mathbf{h}}_\mathrm{B} = [\bar{\mathrm{h}}_2\ \bar{\mathrm{h}}_4]^T$).

A-iii) Antenna elements in a group are highly correlated. That is, $E[|\bar{\mathbf{h}}_\mathrm{A}^H\hat{\mathbf{h}}_r|^2] \approx E[|\bar{\mathbf{h}}_\mathrm{B}^H\hat{\mathbf{h}}_r|^2]$ where $\hat{\mathbf{h}}_r$ is the quantized vector of $\mathbf{h}_r$ and generated from the channel statistic-based codebook.

The following theorem provides an approximate upper bound of the quantization distortion D under these assumptions.

Theorem 3.1: The quantization distortion D of the AGB algorithm under the channel statistic-based codebook satisfies

$$D \lesssim N_t\delta + \eta\sqrt{N_t\left(1-(1-\delta)^2\right)} \tag{20}$$

where $\delta = \frac{\sigma_2^2}{\sigma_1^2} 2^{-\frac{B-B_p}{N_g-1}}$ is an upper bound of the normalized distortion between $\bar{\mathbf{h}}_\mathrm{A}$ and $\hat{\mathbf{h}}_r$ (σ_i is the i-th largest singular value of $\mathbf{R}_{t,\mathrm{A}}$)[3] and η is the correlation coefficient between two random variables $\|\mathbf{h}\|^2$ and $1 - |\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2$.

Proof: Using (16), we have

$$D = E\left[\|\mathbf{h}\|^2\left(1 - |\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right)\right] \tag{21}$$

$$= E\left[\|\mathbf{h}\|^2\right]\left(1 - E\left[|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right]\right) + Cov(\|\mathbf{h}\|^2, 1 - |\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2) \tag{22}$$

$$= E\left[\|\mathbf{h}\|^2\right]\left(1 - E\left[|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right]\right) + \eta\sqrt{Var\left(\|\mathbf{h}\|^2\right)}\sqrt{Var\left(1 - |\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right)} \tag{23}$$

$$\le E\left[\|\mathbf{h}\|^2\right]\left(1 - E\left[|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right]\right) + \eta\sqrt{Var\left[\|\mathbf{h}\|^2\right]}\sqrt{1 - \left(E\left[|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right]\right)^2} \tag{24}$$

where (22) is because[4] $Cov(X,Y) = E[XY] - E[X]E[Y]$,

[3]For example, if $\mathbf{R}_t = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$ and A = {1, 3}, then $\mathbf{R}_{t,\mathrm{A}} = \begin{bmatrix} 1 & \rho^2 \\ \rho^2 & 1 \end{bmatrix}$.

[4]Note that the covariance becomes zero if X and Y are statistically independent. For i.i.d channel, $\|\mathbf{h}_w\|^2$ and $|\bar{\mathbf{h}}_w\mathbf{w}|^2$ are independent where $\bar{\mathbf{h}}_w = \frac{\mathbf{h}_w}{\|\mathbf{h}_w\|}$ and $\mathbf{w}$ is quantized vector [14]. However, due to the spatial correlation, $Cov(\|\mathbf{h}\|^2, 1-|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2)$ cannot be zero in our problem.

TABLE I
SUMMARY OF THE ANTENNA GROUP PATTERN GENERATION.

Initialization	B_p: the number of bits for the pattern set $\mathcal{S}$: the set of patterns to be selected
Main operation	1) Initialize the index set $\Omega = \{1, \ldots, N_{\max}\}$ where $N_{\max} = \frac{\prod_{n=0}^{N_g-1} \binom{N_t - n\kappa}{\kappa}}{N_g!}$. $\kappa = \frac{N_t}{N_g}$ is the number of elements in an antenna group. 2) For each pattern $i \in \Omega$, calculate the Frobenius norm of the quasi-correlation matrix $r_i = \|\tilde{\mathbf{R}}_t^{(i)}\|_F$. Without loss of generality, assume $r_1 \geq r_2 \geq \cdots \geq r_{N_{\max}}$. 3) Choose $L(\geq 2^{B_p})$ patterns $\mathcal{T} = \{r_1, \ldots, r_L\}$. 4) Apply the subspace packing to $\mathcal{T}$ to generate the pattern set $\mathcal{S}$. Construct $N_c = \binom{L}{2^{B_p}}$ candidate sets $\{\mathcal{S}_k\}_{k=1}^{N_c}$ where $\mathcal{S}_k = \{\tilde{\mathbf{R}}_t^{k,1}, \tilde{\mathbf{R}}_t^{k,2}, \cdots, \tilde{\mathbf{R}}_t^{k,2^{B_p}}\}$. $\tilde{\mathbf{R}}_t^{k,i}$ is the i-th quasi-correlation matrix from the k-th candidate set. 5) Calculate the minimum d_{corr} of a $\mathcal{S}_k$ $d_{k,\min}(\mathcal{S}_k) = \min_{1 \leq m \leq n \leq 2^{B_p}} d_{\text{corr}}(\tilde{\mathbf{R}}_t^{k,m}, \tilde{\mathbf{R}}_t^{k,n})$. Decide the pattern set $\mathcal{S}$ $\mathcal{S} = \arg\max_{k=1,\cdots,N_c} d_{k,\min}(\mathcal{S}_k)$.

(23) is because $Cov(X,Y) = \eta\sqrt{Var(X)}\sqrt{Var(Y)}$ and (24) is because $Var\left(1 - |\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right) = E\left[|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^4\right] - \left(E\left[|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right]\right)^2 \leq 1 - \left(E\left[|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right]\right)^2$.

The normalized distortion term $1 - E\left[|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right]$ in the right-hand side of (24) is approximately upper bounded as

$$\begin{aligned}
&1 - E\left[|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2\right] \\
&\overset{(a)}{=} 1 - E\left[|\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r + \bar{\mathbf{h}}_{\mathrm{B}}^H\hat{\mathbf{h}}_r|^2\right] \\
&= 1 - E[|\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r|^2 + |\bar{\mathbf{h}}_{\mathrm{B}}^H\hat{\mathbf{h}}_r|^2 + (\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r)^H(\bar{\mathbf{h}}_{\mathrm{B}}^H\hat{\mathbf{h}}_r) \\
&\quad + (\bar{\mathbf{h}}_{\mathrm{B}}^H\hat{\mathbf{h}}_r)^H(\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r)] \\
&= 1 - E\left[|\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r|^2\right] - E\left[|\bar{\mathbf{h}}_{\mathrm{B}}^H\hat{\mathbf{h}}_r|^2\right] \\
&\quad - 2E\left[\mathrm{Re}(\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r)^H(\bar{\mathbf{h}}_{\mathrm{B}}^H\hat{\mathbf{h}}_r)\right] \\
&\overset{(b)}{\approx} 1 - E\left[|\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r|^2\right] - E\left[|\bar{\mathbf{h}}_{\mathrm{B}}^H\hat{\mathbf{h}}_r|^2\right] \\
&\overset{(c)}{\approx} 1 - 2E\left[|\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r|^2\right] \\
&\overset{(d)}{\leq} 1 - 2(1-\delta)E\left[\|\bar{\mathbf{h}}_{\mathrm{A}}\|^2\right] \\
&\overset{(e)}{=} \delta
\end{aligned} \tag{25}$$

where (a) is because $|\bar{\mathbf{h}}^H\tilde{\mathbf{h}}_r|^2 = |\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r + \bar{\mathbf{h}}_{\mathrm{B}}^H\hat{\mathbf{h}}_r|^2$, and (b) follows from $E[\mathrm{Re}((\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r)^H(\bar{\mathbf{h}}_{\mathrm{B}}^H\hat{\mathbf{h}}_r))] = \sum_i \mathrm{Re}(E[\bar{\mathbf{h}}_{\mathrm{B},i}^H]E[\hat{\mathbf{h}}_r^H\bar{\mathbf{h}}_{\mathrm{A}}\hat{\mathbf{h}}_{r,i}] \approx 0$ (since $E[\bar{\mathbf{h}}_{\mathrm{B},i}^H] = 0$), and (c) follows from A-iii), and (d) follows from the QUB in[5] (18), and (e) follows from $E\left[\|\bar{\mathbf{h}}_{\mathrm{A}}\|^2\right] = \frac{N_g}{N_t} = \frac{1}{2}$.

Plugging (25) into (24), we have

$$\begin{aligned}
D &\leq E\left[\|\mathbf{h}\|^2\right]\delta + \eta\sqrt{Var\left[\|\mathbf{h}\|^2\right]}\sqrt{1-(1-\delta)^2} \\
&= N_t\delta + \eta\sqrt{N_t\left(1-(1-\delta)^2\right)}
\end{aligned}$$

where the last equality follows from $E[\|\mathbf{h}\|^2] = N_t$ and $Var[\|\mathbf{h}\|^2] = N_t$, which is the desired result. ■

[5]Plugging $\bar{\mathbf{h}} = \frac{\bar{\mathbf{h}}_{\mathrm{A}}}{\|\bar{\mathbf{h}}_{\mathrm{A}}\|}$, $\mathcal{C} = \left\{\frac{\mathbf{R}_{t,\mathrm{A}}^{1/2}\mathbf{f}_1}{\|\mathbf{R}_{t,\mathrm{A}}^{1/2}\mathbf{f}_1\|}, \cdots, \frac{\mathbf{R}_{t,\mathrm{A}}^{1/2}\mathbf{f}_{2^{B-B_p}}}{\|\mathbf{R}_{t,\mathrm{A}}^{1/2}\mathbf{f}_{2^{B-B_p}}\|}\right\}$, and $\delta = \frac{\sigma_2^2}{\sigma_1^2}2^{-\frac{B-B_p}{N_g-1}}$ into (18), we get $E\left[|\bar{\mathbf{h}}_{\mathrm{A}}^H\hat{\mathbf{h}}_r|^2\right] \geq (1-\delta)E\left[\|\bar{\mathbf{h}}_{\mathrm{A}}\|^2\right]$.

We note that the relationship between the quantization distortion D and the transmit antenna correlation is not clearly shown in (20). When a specific correlation model is used, however, we can observe the relationship between two. For example, if the exponential correlation model is employed, then the transmit correlation matrix $\mathbf{R}_t$ is expressed as [24]

$$\mathbf{R}_t = \begin{bmatrix} 1 & \rho & \cdots & \rho^{N_t-1} \\ \rho^H & 1 & \cdots & \rho^{N_t-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{(N_t-1)H} & \rho^{(N_t-2)H} & \cdots & 1 \end{bmatrix} \tag{26}$$

where $\rho = \alpha e^{j\theta}$ is the transmit correlation coefficient, and α is the magnitude of correlation coefficient, and θ is the phase of the user. When the number of transmit antennas N_t is large, (non-ordered) singular value μ_i of $\mathbf{R}_t$ becomes approximately [25]

$$\begin{aligned}
\mu_i &\approx \sum_{k=-(N_t-1)}^{N_t-1} \rho^{|k|} e^{j\frac{2\pi ik}{N_t}} \\
&\approx \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\frac{2\pi i}{N_t})}, \quad i = 1, \ldots, N_t.
\end{aligned} \tag{27}$$

Using the first and second largest singular values (i.e., μ_{N_t}, μ_{N_t-1}) of (27), we have

$$\frac{\sigma_2}{\sigma_1} \approx \frac{1+\rho^2-2\rho}{1+\rho^2-2\rho\cos(2\pi\frac{N_t-1}{N_t})}. \tag{28}$$

Noting that $\delta = \frac{\sigma_2^2}{\sigma_1^2}2^{-\frac{B-B_p}{N_g-1}}$ in Theorem 3.1, we have

$$D \lesssim N_t\delta + \eta\sqrt{N_t\left(1-(1-\delta)^2\right)} \tag{29}$$

where

$$\delta \approx \left(\frac{1+\rho^2-2\rho}{1+\rho^2-2\rho\cos(2\pi\frac{N_t-1}{N_t})}\right)^2 2^{-\frac{B-B_p}{N_g-1}}.$$

In (29), we can observe that the quantization distortion D decreases with the correlation coefficient ρ. Fig. 6 plots the

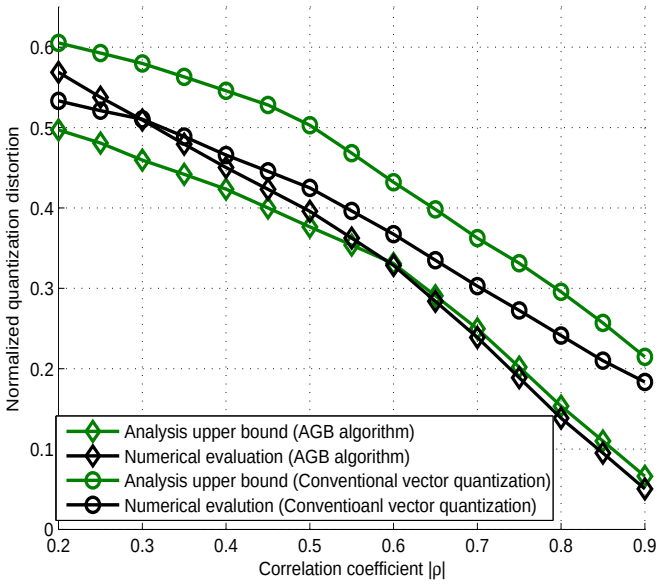


Fig. 6. Normalized quantization distortion as a function of the correlation coefficient ($N_t = 16, N_g = 8, B = 16, B_p = 8$).

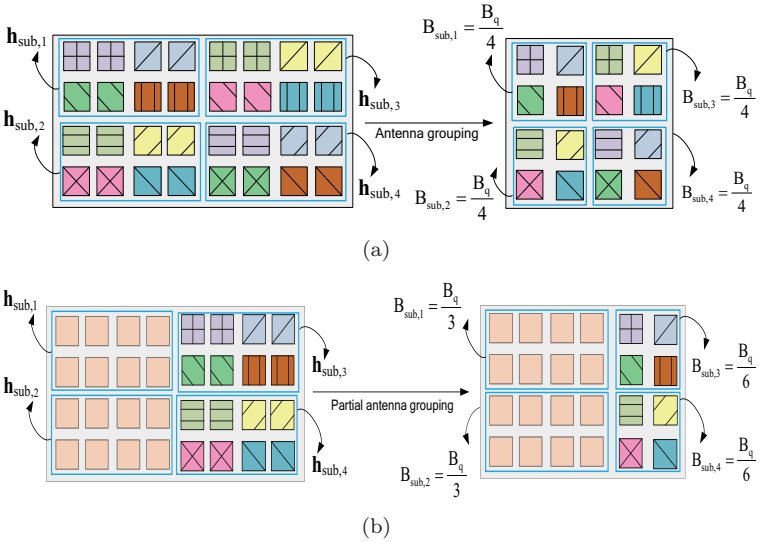


Fig. 7. Example of 4×8 planar array when $N_t = 32, M = 4, M' = 2$ (a) AGB algorithm and (b) Partial antenna grouping scheme.

normalized quantization distortion $\frac{D}{E[\|\mathbf{h}\|^2]}$ as a function of the correlation coefficient ρ. We observe that if $|\rho| > 0.3$, the quantization distortion D of the AGB algorithm is smaller than that of conventional vector quantization. We can also observe that the analysis matches well with the simulation results when the transmit antennas are highly correlated ($|\rho| > 0.6$). However, when the magnitude of ρ is small, the assumption in A-iii) is violated so that the proposed bound is invalid.

D. Partial Antenna Grouping

In this subsection, we discuss a partial antenna grouping scheme that applies the antenna grouping to the part of antenna arrays. The basic idea of this scheme is to partition the antenna array into multiple sub-arrays and then apply the AGB algorithm to the part of sub-arrays having smaller channel gain. In doing so, we can invest more feedback resources to the sub-arrays with higher channel gain. Additionally, the computational complexity required for the pattern generation can be reduced substantially.

Specifically, we first partition the channel vector $\mathbf{h}$ into M sub-arrays $\mathbf{h}_{\text{sub},1}, \cdots, \mathbf{h}_{\text{sub},M}$ ($\mathbf{h}_{\text{sub},i} \in \mathbb{C}^{\frac{N_t}{M}}$). Without loss of generality, we assume that these vectors are ordered based on their channel gain ($\|\mathbf{h}_{\text{sub},1}\|^2 > \cdots > \|\mathbf{h}_{\text{sub},M}\|^2$). Among these, we choose the $M'(\leq M)$ sub-arrays with smallest channel gain $\mathbf{h}_{\text{sub},M-M'+1}, \cdots, \mathbf{h}_{\text{sub},M}$, which we refer to these as weak sub-arrays, and then apply the antenna grouping operation to the weak sub-arrays (see Fig. 7). If we denote the number of groups in the selected sub-arrays as N_g', then the dimension of the channel vector to be quantized becomes $N_g = M'N_g' + (M - M')\frac{N_t}{M}$ where $\frac{N_t}{M}$ is the number of antenna elements for the rest sub-arrays and the number of bit per antenna element becomes $\frac{B_q}{N_g}$ where B_q is the number of bits for vector quantization. Also, the number of bits assigned to the strong and weak sub-arrays are $\frac{N_t}{M}\frac{B_q}{N_g}$ and $N_g'\frac{B_q}{N_g}$, respectively. By setting $\frac{N_t}{M} > N_g'$, we can invest more feedback resources to the strong sub-arrays. As an example, consider the antenna array with 64 elements ($N_t = 64$). If the number of sub-arrays is 4 ($M = 4$) and 2 sub-arrays are chosen as weak sub-arrays ($N_g' = 8$), then $N_g = 48$. Now, suppose the total number of bits for the channel vector quantization is 96 bits, then the proposed approach assigns 32 bits to the strong sub-array and 16 bits to the weak sub-array while the original AGB algorithm assigns 24 bits per each sub-array. By investing 33% more feedback resources into the strong sub-array, quantization loss of the AGB algorithm can be alleviated substantially.

IV. Simulation Results and Discussions

A. Simulation Setup

In this section, we compare the sum rate performance of the conventional vector quantization technique using the channel statistic-based codebook and the proposed AGB algorithm. While all the feedback resources (B-bit) are used to quantize the channel vector $\mathbf{h}_k$ in the conventional approach, B-bit feedback resource is divided into B_q (channel vector quantization), B_p (pattern selection), and B_r (sub-array selection) in the proposed AGB algorithm. To express the feedback allocation, we use the notation $B = (B_q, B_p, B_r)$ in the sequel. As a channel model, we consider the exponential correlation model in (26), one-dimensional uniform linear array (ULA) model [26], and two-dimensional uniform planar array (UPA) model [27]. As a channel vector quantizer, we use RVQ and noncoherent trellis coded quantization (NTCQ). While RVQ is a well-known and standard channel vector quantizer in the literatures, the time to choose the codeword increases

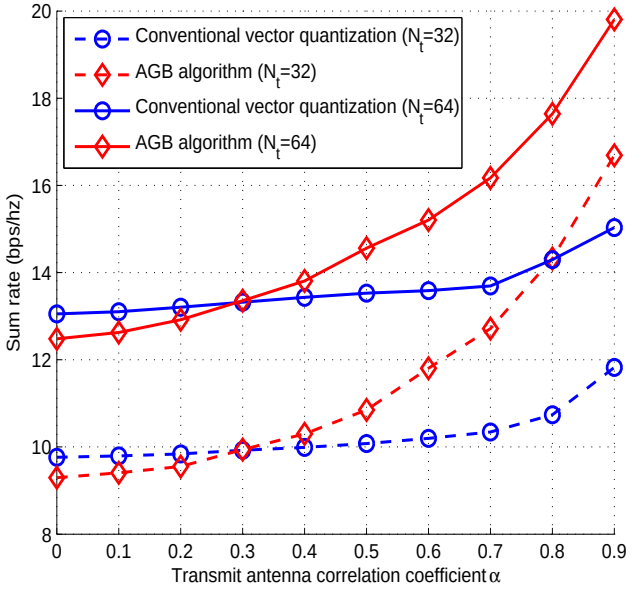


Fig. 8. Sum rate as a function of the transmit antenna correlation coefficient α.

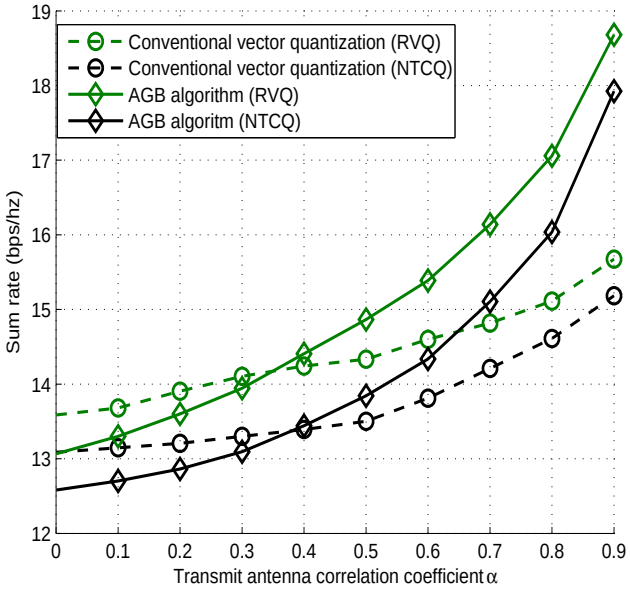


Fig. 9. Sum rate of the NTCQ codebook and analytical approximation of RVQ as a function of α.

exponentially with the quantization bits (when a B-bit codebook is used, we need to check 2^B codewords). Since the number of transmit antennas N_t in our simulations is large (i.e, 32 and 64 antennas), we use the analytical closed-form expression of RVQ [28] instead of simulations. To make sure that the analysis result can be translated into a realizable codebook, we also use the recently proposed low-complexity codebook called NTCQ [10]. Note that the search complexity of the NTCQ codebook scales linearly to the number of transmit antennas so that one can evaluate the performance of NTCQ codebook with reasonable simulation time.

B. Simulation Results

We first consider the exponential channel model [29]

$$r_{ij} = \begin{cases} \rho_k^{|j-i|} & i \leq j \\ (\rho_k^{|j-i|})^H & i > j \end{cases} \tag{30}$$

where r_{ij} is the (i,j)-th element of $\mathbf{R}_{t,k}$ and $\rho_k = \alpha e^{j\theta_k}$ is a transmit correlation coefficient for the k-th user (α is the magnitude of correlation coefficient and θ_k is the phase of the k-th user). Note that the phase of each user is randomly generated from $-\pi$ to π and independent of other users' phases. Note also that all users have the same transmit correlation coefficient $|\rho_k| = \alpha$ since α is determined by the antenna spacing at the basestation. In Fig. 8, we plot the sum rate as a function of α for the multi-user MISO system with $N_t = 32, 64$ and $K = 4$. In our simulations, we allocate one bit per antenna element ($B = N_t$). In the AGB algorithm, we set $B = (17, 14, 1)$ for $N_t = 32$ and $B = (46, 14, 4)$ for $N_t = 64$, respectively. In this case, $N_g = 24$ for $N_t = 32$ and $N_g = 56$ for $N_t = 64$, respectively. When the antenna correlation is low ($\alpha < 0.3$), we observe that the sum rate of the AGB algorithm is slightly worse than that of the conventional vector quantization technique. Whereas, when α increases, the antenna grouping operation becomes effective and thus the sum rate of the AGB algorithm improves drastically. For example, when $\alpha = 0.8$, the sum rate gains of the AGB algorithm over the conventional vector quantization technique is 33% for $N_t = 32$ and 28% for $N_t = 64$, respectively.

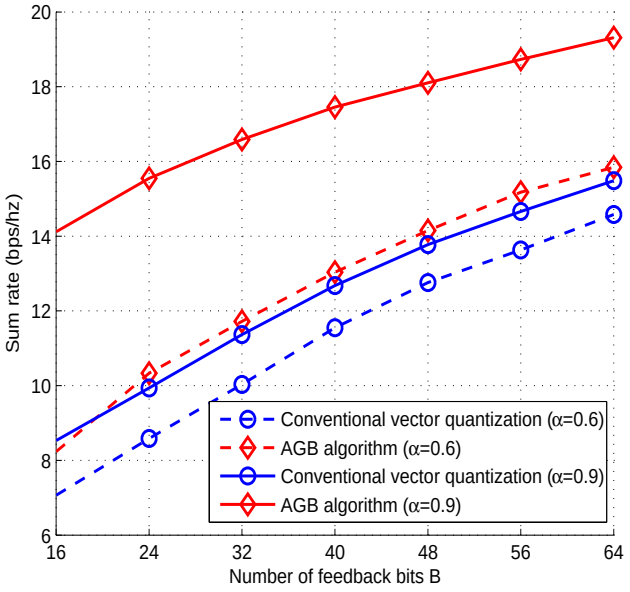


Fig. 10. Sum rate as a function of the number of feedback bits B ($N_t = 32, N_g = 24$).

In order to observe that the analytic result of the RVQ is maintained to the real codebook scenario, we compare the performance between the RVQ and NTCQ codebook. The NTCQ codebook chooses the candidate beamforming vectors using N_g-stage trellis search (each stage selects an entry in each of the candidate vectors). In our simulations, we assign 2 bits per each stage of the trellis search and thus we have $B_q = 2N_g$. For $N_t = 32$ and $N_g = 24$, the feedback allocation is set to $B = (48, 14, 1)$. We observe from Fig. 9 that the overall trend of NTCQ is very similar to that of analytical approximation of RVQ, although there exists a performance gap between two.

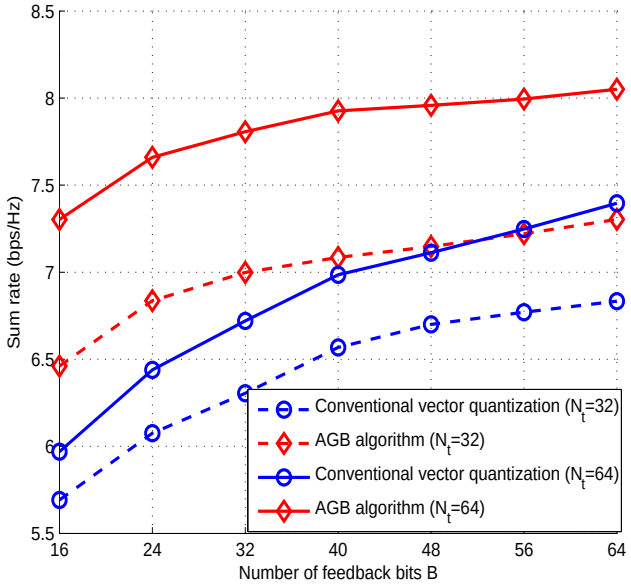
We next measure the sum rate as a function of the number of feedback bits B. In this case, we set $N_t = 32$ and investigate the performance for two scenarios ($\alpha = 0.6$ and 0.9). As shown in Fig. 10, the AGB algorithm achieves significant gain over the conventional vector quantization technique, in particular for the highly correlated scenario ($\alpha = 0.9$). For example, to achieve 14 bps/Hz, 16 and 48 feedback bits are required for the AGB algorithm while 51 and 60 bits are needed for the conventional vector quantization technique, resulting in 68% and 20% feedback overhead reduction for $\alpha = 0.9$ and $\alpha = 0.6$.

We also consider the one-dimensional ULA and two-dimensional UPA ($N_{\mathrm{V}} \times N_{\mathrm{H}}$ array) models, which are more realistic model for massive MIMO scenarios. In the ULA channel model, $\mathbf{R}_{t,k}$ for the k-th user is expressed as

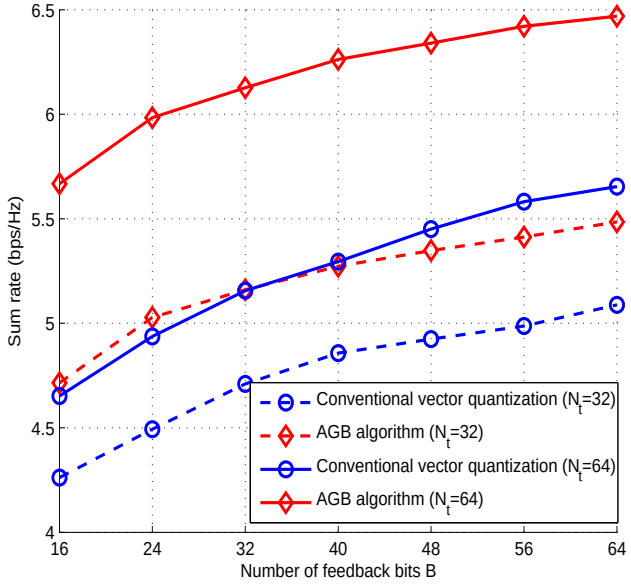
$$[\mathbf{R}_{t,k}]_{m,p} = \frac{1}{2\Delta} \int_{-\Delta+\phi_k}^{\Delta+\phi_k} e^{-j2\pi D(m-p)\sin(\alpha)} d\alpha \qquad (31)$$

where Δ is the angular spread, D is the antenna elements spacing, and ϕ_k is the angle of arrival (AoA) for the k-th user. This ULA model can be extended to the UPA model by combining the vertical correlation matrix $\mathbf{R}_V \in \mathbb{C}^{N_{\mathrm{V}} \times N_{\mathrm{V}}}$ and the horizontal correlation matrix $\mathbf{R}_{H,k} \in \mathbb{C}^{N_{\mathrm{H}} \times N_{\mathrm{H}}}$ using the Kronecker product. The resulting transmit correlation matrix of the UPA model is expressed as $\mathbf{R}_{t,k} = \mathbf{R}_V \otimes \mathbf{R}_{H,k}$ where $\otimes$ is the Kronecker product. We summarize the simulation parameters for the ULA and UPA model in Table II. In Fig. 11, we plot the sum rate as a function of the number of feedback bits for $N_t = 32, 64$ and $K = 1$. For the ULA model, we use $\mathbf{R}_{t,k}$ in (31) for $N_t = 32$ and $N_t = 64$. For the UPA model, we set $N_{\mathrm{V}} = 4, N_{\mathrm{H}} = 8$ for $N_t = 32$ and $N_{\mathrm{V}} = 8, N_{\mathrm{H}} = 8$ for $N_t = 64$, respectively. For both ULA and UPA scenarios, the AGB algorithm outperforms the conventional vector quantization technique with a large margin, resulting in more than 50% feedback overhead reduction.

Finally, in order to observe the effectiveness of the subspace packing approach discussed in Section III.B, we compare the proposed approach and the random pattern generation scheme. In our simulations, we set $B_q = 16, N_t = 16, N_g = 8$ and measure the sum rate as a function of the number of pattern bits B_p. To set the same level of feedback, we set $B = B_q + B_p$ bit for the conventional RVQ based vector quantization. Overall, we observe that the subspace packing approach provides a considerable sum rate gain over the random selection approach as well as the conventional vector quantization technique. For example, to achieve 9 bps/hz, AGB with subspace packing requires $B = 18$ bits while AGB with random patterns and conventional vector quantization require 21 and 24 bits, respectively.



(a)



(b)

Fig. 11. Sum rate as a function of number of feedback bits (a) ULA and (b) UPA correlation model.

V. Conclusions

In this paper, we proposed an efficient feedback reduction algorithm for FDD-based massive MIMO systems. Our work is motivated by the observation that the CSI

TABLE II
SIMULATION PARAMETERS FOR ULA AND UPA MODEL.

Transmit correlation model	Variables	Simulation parameters
Uniform linear array (ULA)	Angular spread	$\Delta = 15^{\circ}$
	Antenna elements spacing	$D = 0.5$
	Angle of arrival for the k-th user	$\phi_k \in (-\pi, \pi]$
Uniform planar array (UPA)	Angular spread (vertical)	$\Delta_V = \frac{1}{2}\left(\arctan(\frac{s+r}{u}) - \arctan(\frac{s-r}{u})\right)$
	Angle of arrival (vertical)	$\phi_V = \frac{1}{2}\left(\arctan(\frac{s+r}{u}) + \arctan(\frac{s-r}{u})\right)$
	Angular spread (horizontal)	$\Delta_H = \arctan(\frac{r}{s})$
	Angle of arrival (horizontal)	$\phi_{H,k} \in (-\pi, \pi]$
	Elevation of the transmit antenna	$u = 60m$
	Radius of the scattering ring for the receiver	$r = 30m$
	Distance from the transmitter	$s = 100m$

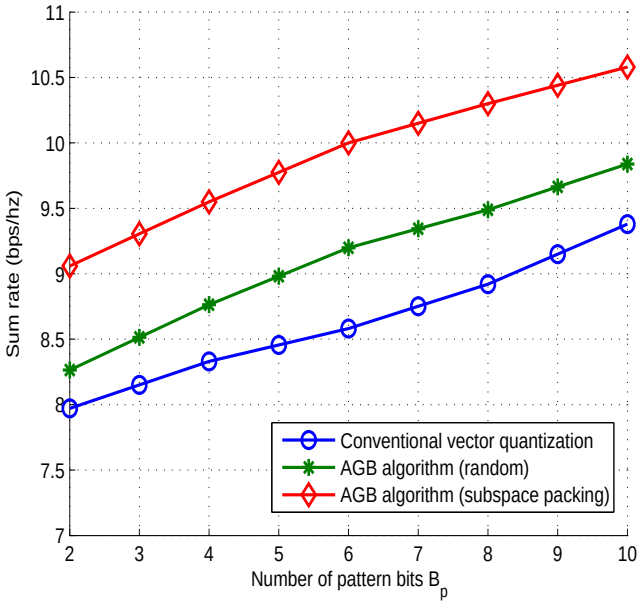


Fig. 12. Sum rate as a function of the number of pattern bits ($N_t = 16, N_g = 8, B_q = 16$).

feedback overhead scales linearly with the number of transmit antennas so that conventional vector approach performing the quantization of the whole channel vector is not an appropriate option for the massive MIMO regime. The key feature of the antenna group beamforming (AGB) algorithm is to control relentless growth of the CSI feedback information in the massive MIMO regime by mapping multiple correlated antenna elements to a single representative value with pre-designed grouping patterns and then choosing the codeword from the codebook generated from the reduced dimension channel vector. It has been shown by distortion analysis and simulation results that the proposed AGB algorithm is effective in achieving a substantial reduction in the feedback overhead in the realistic massive MIMO channels.

Although our study in this work focused on the single-cell scenario, we expect that the effectiveness of the proposed method can be readily extended to multi-cell scenario. In fact, in the multi-cell scenario, more aggressive feedback compression is required since the channel information of the interfering cells as well as the desired cell may be needed at the basestation to properly control inter-cell interference. In this scenario, the proposed AGB algorithm can be used as an effective means to achieve reduction in the feedback information. Also, performance evaluation of the AGB algorithm under channel estimation through closed-loop training [30] and investigation of nonlinear transmitter techniques with user scheduling [31] would be interesting direction to be investigated. Finally, we note that the proposed method can be nicely integrated into the dual codebooks structure in LTE-Advanced [32], [33] by feeding back the pattern index for long-term basis and the codebook index for short-term basis. Since the main target of the massive MIMO system is slowly varying or static channels, dual codebook based AGB algorithm will bring further reduction in feedback overhead.

REFERENCES

[1] B. Lee, J. Choi, D. Love, J. Seol, and B. Shim, "Antenna Grouping based Feedback Reduction Technique for FDD-based Massive MIMO Systems," *Proceedings of IEEE International Conference on Communications*, Jun. 2014.

[2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590-3600, Nov. 2010.

[3] J. Hoydis, S. T. Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?," *IEEE Journal of Selected Areas in Communications,* vol. 31, no. 2, Feb. 2013.

[4] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*, 3GPP TS 36.211 v11.0.0 Std.,

[5] Y. Nam, B. L. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, and J. Lee, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 172-179, Jun. 2013.

[6] D. J. Love, R. W. Heath Jr., V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews., "An overview of limited feedback in wireless communication systems," *IEEE Journal on Selectd Area in Communications*, vol. 26, no. 8, pp. 1341-1365, Oct. 2008.

[7] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2640-2651, Aug. 2011.

[8] E. Björnson, J. Hoydis, M. Kountouris, M. Debbah, "Massive MIMO systems with non-ideal hardware: energy efficiency, estimation, and capacity limits," submitted to *IEEE Transactions on Information Theory*.

[9] C. Shepard, H. Yu, N. Anand, L. E. Li, T. L. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," *Proceedings on ACM Mobile Computing and Networking Conference*, Aug. 2012.

[10] J. Choi, Z. Chance, D. J. Love, and U. Madhow, "Noncoherent trellis coded quantization: A practical limited feedback technique for massive MIMO systems," *IEEE Transactions on Communications*, vol. 61, no. 12, pp. 5016-5029, Dec. 2013.

[11] J. Nam, J. Ahn, A. Adhikary, and G. Caire, "Joint spatial division and multiplexing: Realizing massive MIMO gains with limited channel state information," *Proceedings of Conference on Information Sciences and Systems*, Mar. 2012.

[12] W. Santipach and M. L. Honig, "Asymptotic performance of MIMO wireless channels with limited feedback," *Proceedings on IEEE Military Communications Conference*, Oct. 2003.

[13] N. Jindal, "MIMO broadcast channels with finite rate feedback," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045-5059, Nov. 2006.

[14] C. K. Au-Yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO system," *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 458-462, Feb. 2007.

[15] B. Clerckx, G. Kim, and S. J. Kim, "MU-MIMO with channel statistics based codebooks in spatially correlated channels," *Proceedings on IEEE Global Telecommunications Conference*, Dec. 2008.

[16] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691-1706, Jul. 2003.

[17] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal of Selected Areas in Communications*, vol. 24, no. 3, pp. 528-541, Mar. 2006.

[18] J. Park, B. Lee, and B. Shim, "A MMSE vector precoding with block diagonalization for multiuser MIMO downlink," *IEEE Transactions on Communications*, vol. 60, no. 2, pp. 569-577, Feb. 2012.

[19] P. Ding, D. J. Love, and M. D. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3417-3428, Jul. 2007.

[20] D. J. Love and R. W. Heath Jr., "Limited feedback diversity techniques for correlated channels," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 2, pp. 718-722, Mar. 2006.

[21] D. J. Love, R. W. Heath Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2735-2747, Oct. 2003.

[22] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek, "Correlation matrix diatance, a meaningful measure for evaluation of non-stationary MIMO channels," *Proceedings on IEEE Vehicular Technology Conference*, Jun. 2005.

[23] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and users selection," *IEEE Journal of Selected Areas in Communications*, vol. 25, no. 7, pp. 1478-1491, Sept. 2007.

[24] C. Martin and B. Ottersten, "Asymptotic eigenvalue distributions and capacity for MIMO channels under correlated fading," *IEEE Transactions on Wireless Communications*, vol. 3, no. 4, pp. 1350-1359, Jul. 2004.

[25] R. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Transactions on Information Theory*, vol. 18, no. 6, pp. 725-730, Nov. 1972.

[26] S. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Communications Letters*, vol. 5, no. 9, pp. 369-371, Sep. 2001.

[27] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: The large-scale array regime," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6441-6463, Oct. 2013.

[28] N. Ravindran and N. Jindal, "Multi-user diversity vs. accurate channel state information in MIMO downlink channels," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3037-3046, Sep. 2012.

[29] B. Clerckx, G. Kim and S. J. Kim, "Correlated fading in broadcast MIMO channels: curse or blessing?," *Proceedings on IEEE Global Telecommunications Conference*, Dec. 2008.

[30] J. Choi, D. J. Love, and P. Bidigare, "Downlink training techniques for FDD masive MIMO systems: open-loop and closed-loop training with memory," accepted for *IEEE Journal of Selected Topics in Signal Processing*.

[31] B. Lee and B. Shim, "A vector perturbation with user selection for multiuser MIMO downlink," *IEEE Transactions on Communications*, vol. 60, no. 11, pp. 3322-3331, Nov. 2012.

[32] 3GPP TSG RAN WG1 #62, R1-105011, "Way Forward on 8Tx Codebook for Rel.10 DL MIMO," Aug. 2010.

[33] C. Lim, T. Yoo, B. Clerckx, B. Lee, and B. Shim, "Recent trend of multiuser MIMO in LTE-Advanced," *IEEE Communication Magazine*, vol. 51, no. 3, pp. 127-135, Mar. 2013.