1-5-2022

AoD-Adaptive Channel Feedback for FDD Massive MIMO Systems With Multiple-Antenna Users

Mahmoud Alaaeldin
The American University in Cairo (AUC)

Emad Alsusa

Karim G. Seddik
The American University in Cairo (AUC), KSEDDIK@AUCEGYPT.EDU

Wessam Mesbah

Recommended Citation

APA Citation

Alaaeldin, M. Alsusa, E. Seddik, K. & Mesbah, W. (2022). AoD-Adaptive Channel Feedback for FDD Massive MIMO Systems With Multiple-Antenna Users. *IEEE Access, 10*, 4431–4447. 10.1109/access.2022.3140418
https://fount.aucegypt.edu/faculty_journal_articles/4783

MLA Citation

Alaaeldin, Mahmoud, et al. "AoD-Adaptive Channel Feedback for FDD Massive MIMO Systems With Multiple-Antenna Users." *IEEE Access,* vol. 10, 2022, pp. 4431–4447.
https://fount.aucegypt.edu/faculty_journal_articles/4783

Received December 7, 2021, accepted December 28, 2021, date of publication January 5, 2022, date of current version January 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3140418

AoD-Adaptive Channel Feedback for FDD Massive MIMO Systems With Multiple-Antenna Users

MAHMOUD ALAAELDIN[1], (Member, IEEE), EMAD ALSUSA[1], (Senior Member, IEEE), KARIM G. SEDDIK[2], (Senior Member, IEEE), AND WESSAM MESBAH[3], (Senior Member, IEEE)

[1]Electrical and Electronic Engineering Department, The University of Manchester, Manchester M13 9PL, U.K.
[2]Electronics and Communications Engineering Department, The American University in Cairo, Cairo 11835, Egypt
[3]Electrical Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

Corresponding author: Mahmoud Alaaeldin (mahmoud.alaaeldin@manchester.ac.uk)

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Sklodowska-Curie under Grant 812991.

ABSTRACT In this paper, we propose an efficient feedback scheme for an angle of departure (AoD) based channel estimation in frequency division duplex (FDD) massive multiple-input multiple-output (MIMO) systems with multiple antennas at the users. The channel feedback scheme is based on zero-forcing block diagonalization (BD) and it is proposed for two distinct design cases; in case I, the number of streams intended for a user equals the number of antennas at that user; in case II, the number of streams is less than the number of receive antennas. Case I is applicable in scenarios where high data rate requirements are needed as it transmits data symbols over all of the available degrees of freedom of the system. Diversely, case II is applicable when reliability is a priority in the system as it uses the additional *receive* antennas at the user to achieve spatial diversity to enhance the link performance. The proposed scheme is analyzed for the two cases by quantifying the downlink rate gap from the case of perfect channel state information (CSI). Moreover, we design structured feedback codebooks based on optimal subspace packing in the Grassmannian manifold and show that these codes achieve close performance to the perfect CSI case. Additionally, a vector quantization scheme is proposed to quantize the user's channel matrix when optimal power allocation across multiple streams is adopted in the low signal-to-noise ratio (SNR) region. The feedback codebooks are based on optimal line packing in the Grassmannian manifold, where every vector of the user's channel matrix is quantized and sent to the BaseStation. The results demonstrate a fundamental trade-off between vector quantization, with power optimization across the data streams, and subspace quantization. Specifically, vector quantization codebooks outperform subspace-based codebooks in the low SNR region, while the situation is reversed in the high SNR region.

INDEX TERMS Massive MIMO, FDD, multiple antenna users, block diagonalization, singular values and singular vectors, channel feedback, subspace codebooks, water-filling.

NOMENCLATURE

AoD	Angle of Departure.
BD	Zero-Forcing Block Diagonalization.
BS	BaseStation.
CS	Compressive Sensing.
CSI	Channel State Information.
DFT	Discrete Fourier Transform.
FDD	frequency division duplex
i.i.d.	Independently and Identically Distributed.
LSTM	Long-Short-Term Memory.
MIMO	Multiple-Input Multiple-Output.
mmWave	Millimeter-Wave.
MUSIC	Multiple Signal Classification.
RVQ	Random Vector Quantization.
SNR	Signal to Noise Ratio.
SVD	Singular Value Decomposition.
TDD	Time Division Duplex.
ULA	Uniform Linear Array.
ZF	Zero-Forcing.

The associate editor coordinating the review of this manuscript and approving it for publication was Ibrar Yaqoob.

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) wireless communication systems have been shown to introduce dramatic improvements, in both spectral and energy efficiency, by simultaneously serving multiple users with simple linear precoders [1]–[3]. To fully utilize multiplexing and array gains of massive MIMO, the downlink channel state information (CSI) must be acquired at the BaseStation (BS) to perform precoding and digital beam-forming tasks on the transmitted signals. In time division duplex (TDD) massive MIMO systems, there have been many works in acquiring downlink channels at the BS using the estimated uplink channels by utilizing the channel reciprocity. However, in frequency division duplex (FDD) systems, channel reciprocity cannot be used to obtain the downlink CSI at the BS. In this case, the downlink CSI is generally estimated at the user equipment and then fed back to the BS. The huge number of antennas at the BS leads to an overwhelming overhead, which adversely impacts the system's bandwidth and energy efficiency as well as exacerbating its latency, rendering such a system to be impractical for time-varying channels. Hence reducing this overhead is paramount for realizing the potentials of this technique. FDD deployments provide wider coverage than TDD as mobile equipments in FDD systems transmit on a continuous basis, which enables devices to achieve cell-edge rates farther from the base station [4]. Hence, FDD needs fewer BSs than TDD as long as FDD devices achieve desired cell-edge rates at farther distances. This results in reducing the deployment and operating costs since FDD requires fewer BSs for the same coverage. The previous advantages give a strong motivation to advance FDD massive MIMO topic.

A. RELATED WORK

Several CSI techniques were proposed in the literature for reducing the feedback overhead in FDD massive MIMO systems. For instance, in [5], a spatially common sparsity adaptive channel estimation and feedback scheme was proposed. The authors developed a compressive sensing scheme that exploits the sparse nature of the downlink channels in the angular domain for reliable downlink CSI estimation and feedback with significantly reduced overhead. The authors in [6] also proposed a channel feedback algorithm based on compressive sensing to reduce the feedback load without degrading the quality of channel reconstruction at the BS. They exploited the correlation of CSI to design a quasi-signal-independent dictionary to enhance the quality of CSI recovery at the BS. In [7], the authors exploited the hidden joint sparsity structure in the multi-user MIMO channel matrix by presenting a joint multi-user MIMO channel recovery at the BS. Multiple users fed back distributed channel measurements to the BS to recover the channel matrix using a joint orthogonal matching pursuit algorithm. A robust closed-loop pilot and CSI feedback resource adaptation scheme was proposed in [8], which exploits the joint sparsity of the multi-user massive MIMO channels in order to improve the CSI estimation. Additionally, the framework can minimize the needed pilot and feedback resources for successful CSI recovery. In [9], the authors were able to estimate the users' downlink channel covariance matrix from the uplink pilots using the fact that the angular scattering function of the user channels is invariant over frequency bands. They proposed a novel sparsifying precoder, based on the covariance information, to maximize the effective sparsified channel matrix's rank when the sparsity of each effective user channel is not larger than some desired downlink pilot dimension, resulting in reducing the downlink training and the CSI feedback overhead. In [10], the user first compresses the CSI, based on some compressive sensing approach, then the BS reconstructs the CSI using a deep neural network. A real-time CSI feedback architecture based on a long-short-term memory (LSTM) was proposed in [11] using deep learning CSI sensing and recovery network.

There is also much work on feedback codebook design for massive MIMO systems. For example, a codebook was proposed in [12] for compressed channel feedback for correlated massive MIMO channels. This codebook can quantize and feedback the compressed low-dimensional CSI with reduced overhead. A reduced-dimensional subspace codebook for lens antenna array aided massive MIMO systems was proposed in [13]. By leveraging the concept of angle coherence time, large-dimensional vectors in the channel subspace are first generated. Based on these vectors, the reduced-dimensional subspace codebook is created by considering both the lens and the beam selector. The equivalent channel is quantized and fed back to the BS using this codebook. The authors of [14] proposed a novel dual-stage Grassmannian product quantization approach suitable for high-dimensional CSI. This method works efficiently when the channel can be decomposed in the angular domain, where efficient discrete Fourier transform (DFT) codebooks can be exploited for CSI compression. In [15], the authors proposed a scheme to extrapolate the channel frequency response from the uplink channel estimates to the downlink frequency range. This approach completely removes the need for any feedback from the users' side to the BS. However, the price for this is a degradation in the quality of the downlink channel estimates, due to a downlink spectral efficiency reduction.

In the aforementioned works, the compressive sensing and channel statistics-based feedback techniques suffer from high complexity to provide high-quality channel estimates and reconstruction accuracy at the BS without significantly decreasing the spectral efficiency. In contrast to the above works, an angle of departure (AoD)-adaptive subspace codebook for channel feedback was proposed in [16]. The paper utilized the idea that the angles of departure vary much slower than the channel gains, which results in a significant reduction in the required feedback overhead. This is because the channel vector is constrained in a lower-dimensional subspace of the full M-dimensional space (where M is the number of transmitting antennas at the BS) during the angle

TABLE 1. **Related reviewed works.**

Channel feedback technique	Single Rx antenna	Multiple Rx antennas
CS	[5], [6], [7]	–
DFT	[14]	–
LSTM deep learning	[11]	–
Channel extrapolation	[15]	–
AoD based codebook	[16], [13]	Our proposed design

coherence time. Table 1 categorizes the related reviewed works along with their main distinctive characteristics so as to better position this paper's work by difference.

B. CONTRIBUTIONS

In contrast to existing work, this paper designs and studies an AoD-adaptive channel feedback framework in massive MIMO systems when the users have multiple antennas, which, to the best of our knowledge, has never been addressed before.[1] Different from the above-mentioned channel statistics codebooks, the AoD based massive MIMO channel feedback leverages the concept of angle coherence time in the millimeter-wave (mmWave) channel model. The AoDs information can be estimated with low overhead only once every angle coherence period using the sparse downlink channel estimates at the users. Then, only the low-dimensional channel gains of the resolvable paths of each receive antenna are fed back during the angle coherence period, which we address in this work. As the number of receive antennas increases, the channel feedback overhead increases, which could exhaust the network resources. To overcome this, we propose a feedback scheme that jointly reports the CSI of the receive antennas to the BS resulting in a massive reduction in the required feedback overhead. More specifically, we do not quantize and feed back the channel vector of each receive antenna independently, instead, we quantize and report the subspace spanned by these vectors to the BS which results in a significant rate improvement per each user. To achieve this, we use BD [17] for precoding. Two different design cases are considered; in one, the number of streams intended for a user equals the number of antennas at that user, and in the other case, the number of streams is less than the number of antennas. Since BD involves simultaneous transmissions of multiple data streams to each user while canceling the interference from other users, it only needs the channel subspace of each user's channel matrix at the BS, which requires fewer feedback bits compared to reporting the actual channel matrix. Furthermore, optimally designed subspace codebooks to quantize the subspace of each user's channel matrix are devised. The main goal is to design low overhead feedback schemes while minimizing the system rate loss due to channel quantization. In addition, this paper extends our previous work in [18], [19] by providing detailed mathematical analysis to quantify the rate loss resulting from the proposed *subspace* based quantization scheme for the two considered cases. Moreover, a vector quantization based codebook design, based on water-filling and optimal line packing on the Grassmannian manifold, is proposed for the low signal-to-noise ratio (SNR) region, and it is shown to enhance the downlink spectral efficiency. The contributions of this paper can be summarized as follows:

1) We propose an efficient and structured feedback BD-based AoD-adaptive codebooks using optimal subspace packing on the Grassmannian manifold for massive MIMO systems with multiple antenna users.
2) A channel feedback scheme for two distinct design cases is proposed; in one case the number of streams intended for a user equals the number of antennas at that user, and in the other, the number of streams is less than the number of receive antennas.
3) Detailed performance analysis for the two considered cases is provided to quantify the rate gap between the system with perfect CSI and our proposed scheme, in which we prove that the required number of feedback bits to achieve a constant rate gap scales linearly with SNR.
4) A vector quantization codebook, based on optimal line packing on the Grassmannian manifold, is proposed to enhance the per-user rate in the low SNR region when power allocation (water-filling) across multiple data streams is used, where it is shown that vector quantization codebooks outperform subspace-based codebooks in the low SNR region, while the situation is reversed in the high SNR region.

The rest of the paper is organized as follows. In Section II, the adopted system model is presented, while in Section III, the design of BD-based beam-forming matrices for the two considered scenarios is described. The AoD adaptive subspace codebook used for channel quantization is presented in Section IV. Throughput degradation due to digital channel feedback is analyzed in Section V. The water-filling-based channel quantization and feedback are discussed in Section VI. Simulation results and conclusions are given in Sections VII and VIII, respectively.

Notation: Matrices and vectors are written in boldface letters; matrices are capital letters while vectors are lower-case letters. The transpose and conjugate transpose (Hermitian) of a matrix are denoted by $(\cdot)^T$, $(\cdot)^H$, respectively. $|x|$ is the absolute value of a scalar. The notation $(x)^+$ means that $(x)^+ = 0$ when $x \leq 0$, while $(x)^+ = x$ when $x > 0$. $\mathbb{E}[\cdot]$ denotes the expectation operator. Finally, $\mathbf{I}_P$ denotes the identity matrix of size $P \times P$.

[1] In the multiple receive antenna case, the proposed scheme is based on quantizing and feeding back subspaces not vectors as in the single receive antenna case, where we use Grassmannian subspace packing to design the quantizer. Moreover, to be able to only feed back the channel subspace not the whole matrix, zero-forcing block diagonalization (BD) is used as the interference canceling scheme instead of the regular zero-forcing (ZF) as in the single antenna case. As a consequence, the feedback scheme design is different and all of the derived analytical bounds that evaluate the rate loss of our proposed feedback scheme are different from the single antenna case and provide insights that cannot be drawn from the simpler, single antenna users case.

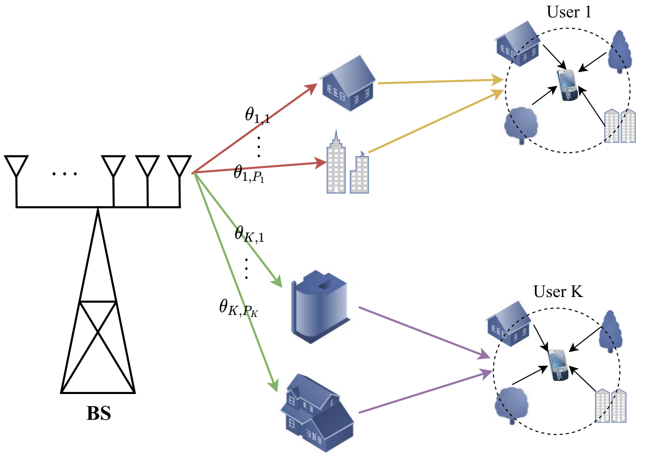


FIGURE 1. The ray-based channel model.

II. SYSTEM MODEL

A. DOWNLINK MASSIVE MIMO CHANNEL MODEL

In this paper, we consider the mmWave massive MIMO broadcast (downlink) system with a single BS communicating with K multi-antenna users as shown in Fig. 1.[2] The BS has M transmitting antennas while the kth user, $\forall k \in \{1, 2, \cdots, K\}$, has N_k receiving antennas. The typical massive MIMO model is used in our system model where the number of transmitting antennas is much higher than the number of users (or users' antennas in our case), i.e., $M \gg \sum_k N_k$ (typically M is in the range of hundreds in massive MIMO). We adopt the narrowband ray-based downlink channel model, which was explained in detail in Chapter 7 of [23] and was adopted in [16] for the case of single receive antenna, for the downlink channels. The number of resolvable paths seen by the BS to the kth user, P_k, depends on the scatters around the BS, which are few typically ranging from 2 to 8, and each path has a different AoD from the BS as shown in Fig. 1. The receive antennas are located in a local multi-path fading environment with many reflectors around it as shown in Fig. 1, hence, the resolvable paths are received at each receive antenna after passing through a multiplicative fading channel. Consequently, each path is multiplied by a complex Rayleigh fading coefficient $g_i \sim \mathcal{CN}(0, 1)$. Therefore, the overall channel vector seen at the jth receive antenna of the kth user is given as

$$\mathbf{h}_{j,k} = \sum_{i=1}^{P_k} g_{j,i}\mathbf{a}(\theta_{k,i}), \tag{1}$$

where parameter $\theta_{k,i}(1 \leq i \leq P_k)$ represents the AoDs of the ith path of the kth user. The transmitting antennas at the BS form a uniform linear array (ULA) [1], where $\mathbf{a}(\theta_{k,i}) \in \mathbb{C}^{1\times M}$ is a steering vector that represents the antenna response of the ith resolvable path of the kth user, and it can be written as

$$\mathbf{a}(\theta_{k,i}) = \left[1, e^{-j2\pi\frac{d}{\lambda}\sin(\theta_{k,i})}, \cdots, e^{-j2\pi\frac{d}{\lambda}(M-1)\sin(\theta_{k,i})}\right], \tag{2}$$

where λ is the signal wavelength, d is the spacing between every two successive antennas at the BS. Although the receive antennas of the same user k see the same AoDs from the BS, they experience different and independent complex path gains. The reason for this is that the receive antennas of the kth user are spatially well-separated so they see completely independent path gains in their local multi-path fading environment. In the mmWave range, it is feasible to have sufficient separations between the few receive antennas as the signal wavelength is already small. Therefore, the channel matrix of the kth user, $\mathbf{H}_k \in \mathbb{C}^{N_k\times M}$, can be expressed as[3]

$$\mathbf{H}_k = \mathbf{G}_k\mathbf{A}_k(\theta_{k,1}, \theta_{k,2}, \cdots, \theta_{k,P_k}), \tag{3}$$

where the matrix $\mathbf{A}_k(\theta_{k,1}, \theta_{k,2}, \cdots, \theta_{k,P_k}) \in \mathbb{C}^{P_k\times M}$ is defined as

$$\mathbf{A}_k(\theta_{k,1}, \theta_{k,2}, \cdots, \theta_{k,P_k}) = \begin{bmatrix} \mathbf{a}(\theta_{k,1}) \\ \mathbf{a}(\theta_{k,2}) \\ \vdots \\ \mathbf{a}(\theta_{k,P_k}) \end{bmatrix}. \tag{4}$$

The row-space of $\mathbf{A}_k$ is called the channel subspace throughout this paper. The jth row of $\mathbf{G}_k \in \mathbb{C}^{N_k\times P_k}$ contains the complex Rayleigh path gains of the jth antenna at the kth user (i.e., the entry $\mathbf{G}_k(j, i)$ represents the complex gain of the ith path of the jth antenna at user k). The complex path gains in $\mathbf{G}_k$ are independently and identically distributed (i.i.d.), as explained earlier, circularly-symmetric complex Gaussian random variables with zero mean and unit variance. Please note that we considered the use of ULAs in this paper for simplicity of presentation. The proposed feedback scheme can work, using the same procedure, with uniform planar arrays (UPA) of antennas with M_1 horizontal antennas and M_2 vertical antennas, where $\mathbf{A}_k(\phi_{k,i}, \theta_{k,i})$ is a function of the azimuth, $\phi_{k,i}$, and the elevation, $\theta_{k,i}$, AoDs of the ith path of the kth user [16]. However, the channel matrix of the kth user is kept the same as $\mathbf{H}_k = \mathbf{G}_k\mathbf{A}_k(\phi_{k,i}, \theta_{k,i})$, hence, the proposed channel feedback technique can also be considered.

[2] The mmWave channel with non line of sight (NLOS) was investigated and compared against the line of sight (LOS) case in [20], where the authors conducted extensive propagation measurement campaigns at 28 GHz and 38 GHz ranges to study the statistical characteristics of mmWave propagation. Other realistic field measurement campaigns, where the NLOS case was extensively studied, at 28 GHz and 73 GHz in New York City (USA) [21], and at $10 - 100$ GHz in Berlin (Germany) [22] were held, and the results were so promising. In light of the previous measurements, the authors in [16] assumed the mmWave channel model in massive MIMO with two-hop scattering with NLOS, where they considered the case of single antenna users.

[3] It should be noted that there will be an angular spread for each scattering cluster; however, the reflected angles within this interval are usually modelled as discrete angles in literature which represent the resolvable paths. Overall, a narrowband clustered channel representation (based on extended Saleh-Valenzuela model) is proposed for mmWave massive MIMO channel, as it allows accurate capturing of the characteristics of mmWave channels. Under this clustered model, the narrowband channel matrix $\mathbf{H}$ is assumed to be the sum of the contributions of P propagation paths [24]–[26]. Based on this, the channel model we adopts in this paper is widely used in the literature, see for example [16], [27]–[31].

The BS sends m_k streams to user k, where $m_k \leq N_k$. Let $\mathbf{d}_k \in \mathbb{C}^{m_k \times 1}$ contain the m_k data symbols to be transmitted simultaneously to the kth user such that

$$\mathbf{d}_k = [u_{k,1} u_{k,2} \cdots u_{k,m_k}]^T. \tag{5}$$

Before transmitting the users' data symbols, the kth user symbol vector is multiplied by the precoding matrix $\mathbf{F}_k \in \mathbb{C}^{M \times m_k}$. Thus, the transmitted vector $\mathbf{x} \in \mathbb{C}^{M \times 1}$, which contains data symbols intended for all the users, is given by

$$\mathbf{x} = \sum_{j=1}^{K} \mathbf{F}_j \mathbf{d}_j, \tag{6}$$

and the received signal at the kth user can be written as

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x} + \mathbf{n}_k = \mathbf{H}_k \mathbf{F}_k \mathbf{d}_k + \mathbf{H}_k \sum_{j=1, j\neq k}^{K} \mathbf{F}_j \mathbf{d}_j + \mathbf{n}_k, \tag{7}$$

where $\mathbf{n}_k \in \mathbb{C}^{N_k \times 1}$ is a circularly symmetric complex Gaussian noise vector at the kth user with a zero vector mean and identity covariance matrix.

The second term in (7) represents the summation of the interference, from the signals intended to all other users in the cell, at user k. The users' precoding matrices, $\mathbf{F}_k$'s, are unitary matrices (i.e., $\mathbf{F}_k^{\mathrm{H}} \mathbf{F}_k = \mathbf{I}_{m_k}$), and in order to adhere to the power constraint, we have $E\left[\|\mathbf{d}_k\|^2\right] = \frac{\gamma}{K}, \forall k \in \{1, 2, \cdots, K\}$, where γ is the total transmit power at the BS.

B. PARTIAL CSI FEEDBACK

In this subsection, we present the CSI feedback elements of the considered channel model that was presented in the previous section. We assume in this paper that each user knows its own downlink CSI.[4] Then, the obtained downlink CSI at each user, $\mathbf{H}_k$, is quantized and fed back to the BS to perform downlink precoding.

The channel matrix, $\mathbf{H}_k$, is composed of two parts; the matrix, $\mathbf{A}_k$, which is a function of the AoDs, and the path gains matrix $\mathbf{G}_k$. We assume throughout the paper that the AoDs are perfectly known at both the user and the BS. This assumption is justified by the fact that the users can estimate the AoDs from downlink channel estimates they already have using the standard multiple signal classification (MUSIC) algorithm [35]. Precisely, the kth user calculates the correlation matrix of the channel vector of the jth receive antenna as $\mathbf{R}_k = \mathbb{E}[\mathbf{h}_{j,k}^{\mathrm{H}} \mathbf{h}_{j,k}]$. The kth user estimates the previous expectation using the sample average technique computed over various channel estimates within the angle coherence time. Therefore, AoDs can be estimated at user k based on $\mathbf{R}_k$ using the traditional MUSIC algorithm. Then, the users quantize and feed back the AoDs, $\theta_{k,i}$, to the BS so it can generate $\mathbf{A}_k$. With reasonably designed number of AoD quantization bits, B_0, the channel sub-spaces, $\mathbf{A}_k$, can fairly assumed to be the same on both sides, without a significant mismatch.

The AoDs are fed back only once every angle coherence time of $\theta_{k,i}$. The coherence time of the AoDs, $\theta_{k,i}$, is relatively very large compared to the coherence time of the path gains in $\mathbf{G}_k$. Consequently, we focus in this paper on studying the quantization and feedback overhead of the path gains matrices, $\mathbf{G}_k$, since it constitutes most of the feedback overhead compared to the relatively insignificant feedback overhead of the AoDs. Therefore, the BS only needs to know the low dimensional path gains matrix $\mathbf{G}_k \in \mathbb{C}^{N_k \times P_k}$ in order to generate the actual channel matrix $\mathbf{H}_k$.

We assume, before Sec. VI, that the power allocated to each user is uniformly divided across its multiple data streams. Hence, in order to perform BD, which will be discussed thoroughly in Sec. III-A, the BS only needs to know the spatial direction of each user's channel matrix. The spatial direction of the kth user, $\widetilde{\mathbf{H}}_k \in \mathbb{C}^{N_k \times M}$, is defined as the row-space of the channel matrix, $\mathbf{H}_k$. The rows of $\widetilde{\mathbf{H}}_k$ are orthonormal and its row-space represents the spatial direction. In the case of massive MIMO channel model described earlier, the spatial direction can be represented as $\widetilde{\mathbf{H}}_k = \frac{1}{\sqrt{M}} \widetilde{\mathbf{G}}_k \mathbf{A}_k$, where the rows of $\widetilde{\mathbf{G}}_k \in \mathbb{C}^{N_k \times P_k}$ are orthonormal and its row-space represents the row-space of $\mathbf{G}_k$. It was proved in [16] that the rows of the channel subspace, $\mathbf{A}_k$, are asymptotically orthogonal, i.e., $\mathbf{A}_k \mathbf{A}_k^{\mathrm{H}} \approx M \mathbf{I}_{P_k}$, as $M \rightarrow \infty$. Since $\mathbf{A}_k$ is known at the BS, $\widetilde{\mathbf{G}}_k$ can be a low-dimensional representative of the spatial direction and its quantization represents the quantization of $\widetilde{\mathbf{H}}_k$. The quantization of $\widetilde{\mathbf{G}}_k$, say $\widehat{\mathbf{G}}_k \in \mathbb{C}^{N_k \times P_k}$, is chosen from the codebook $\mathcal{C}_k = \{\mathbf{C}_{k,1}, \mathbf{C}_{k,2}, \cdots, \mathbf{C}_{k,2^B}\}$, that consists of 2^B low-dimensional quantization sub-spaces in $\mathbb{C}^{N_k \times P_k}$, where B is the number of feedback bits for each user and the rows of $\mathbf{C}_{k,i}$ are orthonormal. The details of the beam-forming matrix design and the codebook design are discussed in Sec. III and Sec. IV, respectively. The kth user quantizes the row-space of its path gains matrix, $\widetilde{\mathbf{G}}_k$, to a quantization subspace, $\widehat{\mathbf{G}}_k = \mathbf{C}_{k,Z_k}$, where the index Z_k is calculated as

$$Z_k = \arg\min_{i \in [1, 2^B]} d^2(\widetilde{\mathbf{G}}_k, \mathbf{C}_{k,i}), \tag{8}$$

where $d(\widetilde{\mathbf{G}}_k, \mathbf{C}_{k,i})$ is the distance metric between the two matrices $\widetilde{\mathbf{G}}_k$ and $\mathbf{C}_{k,i}$. In this paper, we adopt the chordal distance as our distance metric [36], which is given by

$$d(\widetilde{\mathbf{G}}_k, \mathbf{C}_{k,i}) = \sqrt{\sin^2 \phi_1 + \sin^2 \phi_2 + \cdots + \sin^2 \phi_{N_k}}, \tag{9}$$

where the ϕ_j's are the principal angles between the two row-spaces of the matrices $\mathbf{G}_k$ and $\mathbf{C}_{k,i}$ [36]. The rows of each matrix $\mathbf{C}_{k,i} \in \mathcal{C}_k$ are orthonormal (i.e., $\mathbf{C}_{k,i} \mathbf{C}_{k,i}^{\mathrm{H}} = \mathbf{I}_{P_k} \forall\, \mathbf{C}_{k,i} \in \mathcal{C}_k$), and each $\mathbf{C}_{k,i}$ represents a quantization subspace in the codebook. The chordal distance can be calculated

[4] In FDD systems, the downlink channel matrix $\mathbf{H}_k$ is estimated at the user side through downlink channel training. The training overhead for the downlink channel estimation in FDD massive MIMO systems is greatly increased due to the large number of antennas at the BS. However, several effective downlink training methods were proposed to address this problem [5], [27], [32]–[34] by utilizing the angular domain sparsity in the massive MIMO channel model.

as follows

$$d(\widetilde{\mathbf{G}}_k, \mathbf{C}_{k,i}) = \left[N_k - \left\| \widetilde{\mathbf{G}}_k \mathbf{C}_{k,i}^{\mathrm{H}} \right\|_{\mathrm{F}}^2 \right]^{1/2}, \tag{10}$$

where the values of this distance range between 0 and $\sqrt{N_k}$. Note that no channel magnitude feedback is needed at BS in the case of uniform power allocation across the data streams. The BS can hence generate the channel matrix as $\widehat{\mathbf{H}}_k = \frac{1}{\sqrt{M}} \widehat{\mathbf{G}}_k \mathbf{A}_k$.

III. DESIGN OF BD BASED BEAM-FORMING MATRICES

In this section, the procedures of the BD precoding scheme are presented, and the corresponding per-user data rates are stated. The design procedures of the users' beam-forming matrices $\mathbf{F}_k$ are presented for the two considered design cases. Additionally, the corresponding feedback information required for each case will be highlighted.

A. DESIGN OF USERS' BEAM-FORMING MATRICES

BD is considered in this paper as a linear precoding technique that is applied at the BS to serve multiple users in the massive MIMO cell. BD is a precoding technique that completely nulls the interference at each user from other users. In light of the BD procedure, each precoding matrix, $\mathbf{F}_k$, is chosen under the constraint of having $\mathbf{H}_j \mathbf{F}_k = \mathbf{0}, \ \forall j \neq k$. This requires the column space of the precoding matrix to be in the null space of the matrix formed by stacking all $\{\mathbf{H}_j\}_{j \neq k}$ matrices, hence nulling the interference terms in (7) at each user. However, zero interference cannot be achieved practically as the BS does not have perfect knowledge of $\{\mathbf{H}_k\}_{k=1}^K$. In the case of limited feedback, the BS only knows the subspace quantization of the row-space of each $\mathbf{H}_k$, namely, $\widehat{\mathbf{H}}_k$. Then, the BS performs the BD procedure using the quantized subspaces, $\widehat{\mathbf{H}}_1, \widehat{\mathbf{H}}_2, \cdots, \widehat{\mathbf{H}}_K$, to generate the practical precoding matrices, $\widehat{\mathbf{F}}_1, \widehat{\mathbf{F}}_2, \cdots, \widehat{\mathbf{F}}_K$.

The number of receive antennas at user k, N_k, is assumed in this paper to be smaller than the number of resolvable paths P_k, (i.e., $N_k < P_k$). This makes the channel vectors of the receive antennas of user k be linearly independent since they see P_k independent paths with independent path gains (i.e., entries of $\mathbf{G}_k$ are independent). In the following, the two cases we consider when designing the precoding matrices $\widehat{\mathbf{F}}_k$ are presented.

1) CASE I, $N_k = m_k$

In this case, the number of antennas of user k, N_k, is assumed to be equal to the number of data symbols, m_k, transmitted to this user. Define $\mathbf{W}_k$ as

$$\mathbf{W}_k = \left[\widehat{\mathbf{H}}_1^T \cdots \widehat{\mathbf{H}}_{k-1}^T \widehat{\mathbf{H}}_{k+1}^T \cdots \widehat{\mathbf{H}}_K^T \right]^T, \tag{11}$$

where $\widehat{\mathbf{H}}_k, \ k \in \{1, 2, \cdots, K\}$, is the quantized feedback version of the original spatial direction $\widetilde{\mathbf{H}}_k$ of the kth user. The precoding matrix, $\widehat{\mathbf{F}}_k$, of the kth user is forced to lie in the null space of $\mathbf{W}_k$ to achieve the zero interference constraint. Since the AoDs differ from one user to another, the channel subspaces of the users, $\mathbf{A}_k$, are also independent from each other as they depend only on the AoDs of the users. Then, we can infer that the spatial directions of the users, $\widehat{\mathbf{H}}_k$, are linearly independent as well as they lie in the channel subspaces. The rank of $\mathbf{W}_k$ is $\widetilde{L}_k = \mathrm{rank}(\mathbf{W}_k) = N_R - N_k$, where N_R is the total number of receive antennas and $M \gg N_R$. Define the singular value decomposition (SVD) of $\mathbf{W}_k$ as

$$\mathbf{W}_k = \mathbf{U}_k \Sigma_k \left[\mathbf{V}_k^{(1)} \ \mathbf{V}_k^{(0)} \right]^{\mathbf{H}}, \tag{12}$$

where $\mathbf{V}_k^{(1)}$ holds the first $\widetilde{L}_k$ right singular vectors, while $\mathbf{V}_k^{(0)}$ have the remaining $(M - \widetilde{L}_k)$ right singular vectors. Hence, $\mathbf{V}_k^{(0)}$ forms an orthonormal basis for the null space of $\mathbf{W}_k$, and therefore, its columns are candidates for the columns of the kth user precoding matrix, $\widehat{\mathbf{F}}_k$.

The effective channel of the kth user is the product, $\widehat{\mathbf{H}}_k \mathbf{V}_k^{(0)}$. As long as interference from other users is canceled, the precoder selection problem now is equivalent to the single-user MIMO capacity maximization problem. Consequently, the best precoder is the right singular vectors of that effective channel [37]. The rank of the effective channel, $\widehat{\mathbf{H}}_k \mathbf{V}_k^{(0)}$, is $\bar{L}_k$, and it is upper bounded as $\bar{L}_k \leq min\{L_k, \widetilde{L}_k\}$, where L_k is the rank of the quantized channel, $\widehat{\mathbf{H}}_k$. Hence, the SVD of the effective channel of user k is given as

$$\widehat{\mathbf{H}}_k \mathbf{V}_k^{(0)} = \mathbf{Q}_k \begin{bmatrix} \Lambda_k & 0 \\ 0 & 0 \end{bmatrix} \left[\mathbf{R}_k^{(1)} \ \mathbf{R}_k^{(0)} \right]^{\mathrm{H}}, \tag{13}$$

where Λ_k is $\bar{L}_k \times \bar{L}_k$ and the columns of $\mathbf{R}_k^{(1)}$ are the first $\bar{L}_k$ singular vectors. Finally, the product $\mathbf{V}_k^{(0)} \mathbf{R}_k^{(1)}$ forms an orthonormal basis of dimension $\bar{L}_k$, and it represents the precoding matrix that maximizes the capacity of the kth user while achieving zero interference. Hence, the precoding matrix is written as

$$\widehat{\mathbf{F}}_k = \mathbf{V}_k^{(0)} \mathbf{R}_k^{(1)}. \tag{14}$$

2) CASE II, $N_k > m_k$

In this case, the number of antennas at the kth user, N_k, is assumed to be greater than the number of data symbols, m_k, transmitted to that user. Adding more receive antennas enhances the diversity gain at the users. Since the number of receive antennas at the user k is greater than the number of data streams intended to it, feeding back the whole spatial direction, $\widetilde{\mathbf{H}}_k \in \mathbb{C}^{N_k \times M}$, as in case I, is not needed. For case II, only the subspace spanned by the first m_k right singular vectors of the channel matrix $\mathbf{H}_k$ is needed to be fed back to the BS. Let the SVD of the channel matrix $\mathbf{H}_k$ of the kth user be

$$\mathbf{H}_k = \mathbf{U}_k \Sigma_k \mathbf{V}_k^{\mathrm{H}}, \tag{15}$$

where $\mathbf{U}_k \in \mathbb{C}^{N_k \times N_k}$ and $\mathbf{V}_k \in \mathbb{C}^{M \times M}$ are unitary matrices, and $\Sigma_k \in \mathbb{C}^{N_k \times M}$ is a rectangular matrix that has the singular values on its diagonal. Let $\mathbf{V}_{k,m_k}$ be a matrix that contains the first m_k columns of $\mathbf{V}_k$. The column-space of $\mathbf{V}_{k,m_k}$ needs to be quantized and fed back. Taking the Hermitian operator on both sides of (15), we get

$$\mathbf{H}_k^{\mathrm{H}} = \mathbf{V}_k \Sigma_k^{\mathrm{H}} \mathbf{U}_k^{\mathrm{H}}. \tag{16}$$

From (16), as long as the off-diagonal elements of Σ_k^{H} are all zeros, we can notice that each column of $\mathbf{H}_k^{\mathrm{H}}$ is a linear combination of the first N_k columns of $\mathbf{V}_k$. Thus, the subspace spanned by the first N_k columns of $\mathbf{V}_k$ is equivalent to column space of $\mathbf{H}_k^{\mathrm{H}} \in \mathbb{C}^{M\times N_k}$. Consequently, we can conclude that the column-space of $\mathbf{V}_{k,m_k}$ always lies in the column-space of $\mathbf{H}_k^{\mathrm{H}}$ as $m_k < N_k$. As long as the column space of $\mathbf{H}_k^{\mathrm{H}}$ always lies in the column-space of $\mathbf{A}_k^{\mathrm{H}}$, then the column-space of $\mathbf{V}_{k,m_k}$ always lies in the column-space of $\mathbf{A}_k^{\mathrm{H}} \in \mathbb{C}^{M\times P_k}$ too. This is important since $\mathbf{A}_k$ is assumed to be already known at the BS. Therefore, a low-dimensional codebook, to be designed in Sec. IV, can be used to quantize $\mathbf{V}_{k,m_k}$. In [38], the columns of $\mathbf{V}_{k,m_k}$ was proved to be isotropically distributed on the subspace in which they lie. Consequently, a Grassmannian subspace packing based codebook can be used to quantize $\mathbf{V}_{k,m_k}$, which will be presented in Sec. IV-B. Let the quantized version of $\mathbf{V}_{k,m_k}$ be $\widehat{\mathbf{V}}_{k,m_k} \in \mathbb{C}^{M\times m_k}$.

Now, let $\mathbf{S}_k \in \mathbb{C}^{M\times(M-\sum_{i=1,i\neq k}^{K} m_i)}$ represent the orthonormal basis of the null space of $\mathbf{W}_k$, where

$$\mathbf{W}_k = \left[\widehat{\mathbf{V}}_{1,m_1} \cdots \widehat{\mathbf{V}}_{k-1,m_{k-1}} \widehat{\mathbf{V}}_{k+1,m_{k+1}} \cdots \widehat{\mathbf{V}}_{K,m_K}\right]^{\mathrm{H}}. \tag{17}$$

The effective channel of the kth user will be the product $\widehat{\mathbf{V}}_{k,m_k}^{\mathrm{H}}\mathbf{S}_k$. The SVD of this product is given by

$$\widehat{\mathbf{V}}_{k,m_k}^{\mathrm{H}}\mathbf{S}_k = \mathbf{Q}_k \Lambda_k \left[\mathbf{R}_k^{(1)}\ \ \mathbf{R}_k^{(0)}\right]^{\mathbf{H}}, \tag{18}$$

where $\mathbf{R}_k^{(1)}$ represents the first m_k right singular vectors. Finally, the product, $\mathbf{S}_k\mathbf{R}_k^{(1)}$, forms an orthonormal basis of dimension m_k, and it represents the precoding matrix $\widehat{\mathbf{F}}_k \in \mathbb{C}^{M\times m_k}$ that maximizes the capacity of the kth user while achieving zero interference. The precoding matrix $\widehat{\mathbf{F}}_k$ is given by

$$\widehat{\mathbf{F}}_k = \mathbf{S}_k\mathbf{R}_k^{(1)}. \tag{19}$$

Hence, the received vector $\mathbf{y}_k \in \mathbb{C}^{N_k\times 1}$ at user k becomes

$$\mathbf{y}_k = \mathbf{H}_k\widehat{\mathbf{F}}_k\mathbf{d}_k + \sum_{j=1,j\neq k}^{K} \mathbf{H}_k\widehat{\mathbf{F}}_j\mathbf{d}_j + \mathbf{n}_k. \tag{20}$$

The received vector $\mathbf{y}_k$, in (20), is finally left multiplied by $\mathbf{U}_{k,m_k}^{\mathrm{H}}$, where $\mathbf{U}_{k,m_k} \in \mathbb{C}^{N_k\times m_k}$ is the matrix that contains the first m_k columns of the matrix $\mathbf{U}_k$ given in (15).

B. THE PER-USER RATE

The BS uses the quantized feedback, $\widehat{\mathbf{H}}_k$, to compute the precoding matrices, $\widehat{\mathbf{F}}_k$, and perform downlink precoding on the transmitted data vector of user k, $\mathbf{d}_k \in \mathbb{C}^{m_k\times 1}$. BD is a linear precoding scheme that cancels the interference at each user due to all other users as discussed in Sec. III-A. Therefore, the interference term in (7) is completely canceled in the case of perfect CSI knowledge at the BS, i.e., $\widehat{\mathbf{H}}_k \equiv \widetilde{\mathbf{H}}_k$. Then, the per-user ergodic rate for both cases I & II is given by [17], [38]

$$R_{\mathrm{CSIT}}(\gamma) = \mathbb{E}\log_2\left|\mathbf{I}_N + \frac{\gamma}{Km_k}\mathbf{H}_k\mathbf{F}_k\mathbf{F}_k^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}\right|, \tag{21}$$

where uniform power allocation across the multiple data streams is adopted.

In practical systems, the BS cannot have ideal downlink CSI information because the feedback is limited with B bits per user. Consequently, the second term in (7), which represents the interference at user k due to all other users, cannot be totally canceled because the row-space of $\widehat{\mathbf{H}}_k$ is not exactly the same as the true spatial direction of the channel, i.e., the row space of $\widetilde{\mathbf{H}}_k$. Hence, some residual interference power will remain due to subspace quantization of the channel, and hence the per-user rate for both cases I & II is given as [39]

$$\begin{aligned} R_{\mathrm{QUANT}}(\gamma) = \mathbb{E}\log_2\left|\mathbf{I}_N + \frac{\gamma}{Km_k}\sum_{j=1}^{K}\mathbf{H}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}\right| \\ -\mathbb{E}\log_2\left|\mathbf{I}_N + \frac{\gamma}{Km_k}\sum_{j=1,j\neq k}^{K}\mathbf{H}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}\right|, \end{aligned} \tag{22}$$

where the term $\mathbf{H}_k\widehat{\mathbf{F}}_k\widehat{\mathbf{F}}_k^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}$ represents the useful signal intended for user k and, $\sum_{j=1,j\neq k}^{K}\mathbf{H}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}$ represents the multi-user interference at user k. It should be noted that (21) and (22) are optimistic ergodic rates since we assume that the receiver has perfect knowledge of the combined channel matrix, $\mathbf{H}_k\mathbf{F}_k$, and also of the whole interference covariance matrix. This can be approached by pilot schemes in the downlink streams.

IV. AoD-ADAPTIVE SUBSPACE CODEBOOK

The AoDs of the ith path of user k, $\theta_{k,i}$, which were defined in (3), depend on the obstacles that surround the BS. These obstacles change their physical positions very slowly compared with the channel path gains coherence time. On the other hand, the path gains in $\mathbf{G}_k$ depend on the scatterers that surround user k. Therefore, the path gains in $\mathbf{G}_k$, changes much faster than the path AoDs, $\theta_{k,i}$s, [16], [40]. Hence, the coherence time of the AoDs of a resolvable path is much longer than the coherence time of its path gains. However, the size of $\mathbf{G}_k$ is much lower than the size of the channel matrix, $\mathbf{H}_k$, which substantially reduces the feedback overhead. The spatial direction of the user k, $\widetilde{\mathbf{H}}_k$, is asymptotically isotropically distributed in its channel subspace, i.e., the row-space of $\mathbf{A}_k(\theta_{k,1},\cdots,\theta_{k,P_k})$, during the angle coherence time. As in (3), we can see that each row of the channel matrix, $\mathbf{H}_k$, is a linear combination of the P_k resolvable paths of user k, where $\mathbf{A}_k$ is determined by the AoDs. $\widetilde{\mathbf{H}}_k$ is asymptotically isotropically distributed in the row-space of $\mathbf{A}_k$ because the rows of $\mathbf{A}_k$, i.e., steering vectors, are asymptotically orthogonal to each other, i.e., $\mathbf{A}_k\mathbf{A}_k^{\mathrm{H}} \approx M\mathbf{I}_{P_k}$ [16],

Algorithm 1: Feedback Summary

```
Input: $\mathbf{A}_k$, $\mathbf{G}_k$;
if Case I then
    Quantize $\mathbf{G}_k$ to $\widehat{\mathbf{G}}_k = \mathbf{C}_{k,Z_k}$, where $Z_k$ is calculated as in (8);
    The index $Z_k$ is fed back to the BS;
end
if Case II then
    Calculate $\mathbf{V}_{k,m_k}$ as in Sec. III-A2;
    Decompose $\mathbf{V}_{k,m_k}$ as $\mathbf{V}_{k,m_k} = \frac{1}{\sqrt{M}}\mathbf{A}_k^{\mathrm{H}}\mathbf{J}_k$;
    Quantize $\mathbf{J}_k$ to $\widehat{\mathbf{J}}_k = \mathbf{C}_{\mathrm{II},k,Z_{\mathrm{II},k}}$, where $Z_{\mathrm{II},k}$ is calculated as in (23);
    The index $Z_{\mathrm{II},k}$ is fed back to the BS;
end
```

and the path gains in $\mathbf{G}_k$ are modeled as i.i.d. circularly symmetric complex Gaussian random variables with zero mean and unit variance. This makes the spatial direction of each user to be asymptotically uniformly distributed in its channel subspace during the angle coherence time, and this justifies the use of subspace packing based channel quantizer to quantize and feedback the channel matrices. The number of paths, P_k, is greatly less than the number of transmit antennas, M, at the BS because of the limited scattering of mmWave [20]. Hence, the row-space of $\mathbf{A}_k$ is a low-dimensional subspace of the full M-dimensional ambient space. As long as the BS knows the AoDs, only the row-space of the path gains matrix, $\mathbf{G}_k \in \mathbb{C}^{N_k \times P_k}$, needs to be quantized and fed back to the BS. Then, for case I, $\widehat{\mathbf{G}}_k \in \mathbb{C}^{N_k \times P_k}$, is chosen from the codebook $\mathcal{C}_k = \{\mathbf{C}_{k,1}, \mathbf{C}_{k,2}, \cdots, \mathbf{C}_{k,2^B}\}$, that consists of 2^B low-dimensional quantization sub-spaces in $\mathbb{C}^{N_k \times P_k}$. $\widehat{\mathbf{G}}_k$ is chosen from the codebook according to (8). The rows of $\widehat{\mathbf{G}}_k$ are orthonormal, and its row-space is isotropically distributed over the complex P_k-dimensional space.

For case II, since the column-space of $\mathbf{V}_{k,m_k}$ lies in the column-space of $\mathbf{A}_k^{\mathrm{H}}$, as proved in Sec. III-A2, then $\mathbf{V}_{k,m_k}$ can be represented as $\mathbf{V}_{k,m_k} = \frac{1}{\sqrt{M}}\mathbf{A}_k^{\mathrm{H}}\mathbf{J}_k$, where $\mathbf{J}_k \in \mathbb{C}^{P_k \times m_k}$ is a unitary matrix. As long as the BS knows $\mathbf{A}_k$, we can only quantize and feed back the column space of the low-dimensional matrix $\mathbf{J}_k$. Hence, a low-dimensional codebook $\mathcal{C}_{\mathrm{II},k} = \{\mathbf{C}_{\mathrm{II},k,1}, \mathbf{C}_{\mathrm{II},k,2}, \cdots, \mathbf{C}_{\mathrm{II},k,2^B}\}$, that consists of 2^B low-dimensional quantization sub-spaces in $\mathbb{C}^{P_k \times m_k}$, can be used to quantize $\mathbf{J}_k$ in case II. The matrices $\mathbf{C}_{\mathrm{II},k,i} \in \mathbb{C}^{P_k \times m_k}$ are unitary and represent the quantization subspaces of case II AoD-adaptive subspace codebook. Then, $\widehat{\mathbf{J}}_k = \mathbf{C}_{\mathrm{II},k,Z_{\mathrm{II},k}} \in \mathbb{C}^{P_k \times m_k}$ is chosen from $\mathcal{C}_{\mathrm{II},k}$ according to

$$Z_{\mathrm{II},k} = \underset{i \in [1,2^B]}{\arg\min} \left[m_k - \left\| \mathbf{J}_k^{\mathrm{H}} \mathbf{C}_{\mathrm{II},k,i} \right\|_{\mathrm{F}}^2 \right]. \tag{23}$$

The column-space of $\widehat{\mathbf{J}}_k$ is isotropically distributed over the complex P_k-dimensional space. Algorithm 1 summarizes the feedback procedures for both cases I & II.

A. RANDOM SUBSPACE QUANTIZATION CODEBOOKS

Generally, obtaining optimal quantization codebooks is not an easy task, especially when the number of quantization subspaces is large. Therefore, the performance of such codebooks can be studied by averaging over random codebooks [41]. Analyzing the performance of random codebooks is much easier, providing us with some useful performance bounds for structured codes. In our subspace quantization problem, a set of 2^B m_k-dimensional subspaces is randomly picked in the ambient P_k-dimensional Euclidean space. The infinite set containing all subspaces of dimension m_k in the Euclidean P_k-dimensional space is called Grassmannian manifold, which is denoted by $\mathcal{G}_{P_k,m_k}$. The 2^B random quantization subspaces in our random quantization codebook are uniformly distributed over the Grassmannian manifold $\mathcal{G}_{P_k,m_k}$. We can pick a quantization subspace over over $\mathcal{G}_{P_k,m_k}$ at random by generating a matrix of dimension $m_k \times P_k$ whose all entries are i.i.d. complex Gaussian. Then, using QR decomposition, an orthonormal basis for the row-space of the generated random matrix is obtained to represent the randomly picked quantization subspace. Then, the average quantization error can be calculated by averaging over many random codebooks. However, random codebooks cannot be used in practical systems.

B. GRASSMANNIAN SUBSPACE PACKING

As long as the row and column spaces of $\widehat{\mathbf{G}}_k$ and $\widehat{\mathbf{J}}_k$ respectively are isotropically distributed over $\mathcal{G}_{P_k,m_k}$, we can solve a subspace packing problem in the Grassmannian manifold to obtain a structured codebook to be used in practical systems. The subspace packing problem is defined by finding 2^B subspaces in a higher dimensional space where the minimum distance between any two subspaces in the set is maximized. The chordal distance, defined in (10), is used in this paper as the distance metric between the subspaces. By solving the subspace packing problem and finding a good set of 2^B quantization subspaces, we can construct a Grassmannian subspace codebook. An iterative algorithm in [36] is used to solve the subspace packing problem. When the number of subspaces in the codebook, 2^B, is lower than P_k^2, the minimum inter-distance between the subspaces in the Grassmannian codebook reaches an upper bound called the Rankin bound [36]. This upper bound is the maximum attainable theoretical distance that can be achieved.

V. THROUGHPUT ANALYSIS

In this section, we calculate the rate gap between the ideal rate and the rate using a random subspace quantization scheme. The rate gap is calculated assuming that all users have the same number of receive antennas, i.e., $N_k = N$, the same number of data streams, i.e., $m_k = m$, and the same number of resolvable paths, i.e., $P_k = P$. Then, an expression for the required number of feedback bits to achieve some constant rate gap is derived, where we prove that the number of bits scales linearly with the transmit power γ_{dB} in dB.

A. RATE GAP CALCULATIONS

The per-user rate of the ideal CSI case is given by (21), and the per-user rate of the quantized CSI case is given by (22). Both cases I & II follow the same analysis in calculating the rate gap because the channel feedback in both cases, $\widehat{\mathbf{H}}_k$ and $\widehat{\mathbf{V}}_k$, has the same isotropic distribution [38] in the channel subspace $\mathbf{A}_k$. Then, the same BD procedure is applied in both cases to calculate the precoding matrices. Following Theorem 1 of [39], which gives an upper bound for the rate gap in Multi-User MIMO systems, we derive an expression for the per-user rate gap due to limited feedback in our massive MIMO system model.

The per-user rate gap $\Delta R(\gamma)$ is bounded as follows

$$\Delta R(\gamma) = R_{\mathrm{CSIT}}(\gamma) - R_{\mathrm{QUANT}}(\gamma) \tag{24}$$

$$\begin{aligned} &\overset{(a)}{\leq} \mathbb{E}\log_2\left|\mathbf{I}_N + \frac{\gamma}{Km}\mathbf{H}_k\mathbf{F}_k\mathbf{F}_k^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}\right| \\ &\quad -\mathbb{E}\log_2\left|\mathbf{I}_N + \frac{\gamma}{Km}\mathbf{H}_k\widehat{\mathbf{F}}_k\widehat{\mathbf{F}}_k^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}\right| \\ &\quad +\mathbb{E}\log_2\left|\mathbf{I}_N + \frac{\gamma}{Km}\sum_{\substack{j=1\\ j\neq k}}^{K}\mathbf{H}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}\right| \end{aligned} \tag{25}$$

$$\overset{(b)}{=} \mathbb{E}\log_2\left|\mathbf{I}_N + \frac{\gamma}{Km}\sum_{\substack{j=1\\ j\neq k}}^{K}\mathbf{H}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}\right| \tag{26}$$

$$\overset{(c)}{=} \mathbb{E}\log_2\left|\mathbf{I}_N + \frac{\gamma}{Km}\widetilde{\mathbf{H}}_k\left(\sum_{j\neq k}\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\right)\widetilde{\mathbf{H}}_k^{\mathrm{H}}\beta_k\right| \tag{27}$$

$$\overset{(d)}{\leq} \log_2\left|\mathbf{I}_N + \frac{\gamma MPN}{Km}(K-1)\mathbb{E}\left[\widetilde{\mathbf{H}}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\widetilde{\mathbf{H}}_k^{\mathrm{H}}\right]\right| \tag{28}$$

Here, (a) follows by neglecting the positive semi-definite interference terms in the quantity

$$\mathbb{E}\log_2\left|\mathbf{I}_N + \frac{\gamma}{Km}\sum_{j=1}^{K}\mathbf{H}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}\right|. \tag{29}$$

Following the BD procedure, both $\mathbf{F}_k$ and $\widehat{\mathbf{F}}_k$ are asymptotically isotropically distributed, and they are chosen independent of $\mathbf{H}_k$. This means that the first two terms in (25) are the same and hence gives (b). By writing $\mathbf{H}_k^H\mathbf{H}_k = \widetilde{\mathbf{H}}_k^H\beta_k\widetilde{\mathbf{H}}_k$, where the rows of $\widetilde{\mathbf{H}}_k \in \mathbb{C}^{N\times M}$ forms an orthonormal basis for the subspace spanned by the rows of $\mathbf{H}_k$, and β_k are the N non-zero and unordered eigenvalues of $\mathbf{H}_k^H\mathbf{H}_k$. Step (c) follows using the fact that $|\mathbf{I}+\mathbf{AB}| = |\mathbf{I}+\mathbf{BA}|$ for matrices $\mathbf{A}$ and $\mathbf{B}$. Finally, (d) follows from Jensen's inequality due to the concavity of log, noting that $\mathbb{E}[\beta_k] = MPN\mathbf{I}_N$ [39].

The value $\mathbb{E}\left[\widetilde{\mathbf{H}}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\widetilde{\mathbf{H}}_k^{\mathrm{H}}\right]$ is computed as follows. First, the channel subspace $\widetilde{\mathbf{H}}_k$ can be decomposed as lemma 1 in [39] as follows

$$\widetilde{\mathbf{H}}_k = \mathbf{Y}_k\mathbf{T}_k\widehat{\mathbf{H}}_k + \mathbf{Z}_k\mathbf{M}_k, \tag{30}$$

where $\mathbf{T}_k \in \mathbb{C}^{N\times N}$ is unitary and uniformly distributed over $\mathcal{G}_{N,N}$, $\mathbf{Z}_k \in \mathbb{C}^{N\times N}$ is lower triangular with positive diagonal elements and satisfies $\mathrm{tr}(\mathbf{Z}_k\mathbf{Z}_k^{\mathrm{H}}) = d^2(\mathbf{H}_k, \widehat{\mathbf{H}}_k)$, $\mathbf{Y}_k \in \mathbb{C}^{N\times N}$ is lower triangular with positive diagonal elements satisfying $\mathbf{Y}_k\mathbf{Y}_k^{\mathrm{H}} = \mathbf{I}_N - \mathbf{Z}_k\mathbf{Z}_k^{\mathrm{H}}$, and the rows of $\mathbf{M}_k \in \mathbb{C}^{N\times M}$ form an orthonormal basis for an isotropically distributed complex N dimensional subspace in the $M-N$ dimensional right nullspace of $\widehat{\mathbf{H}}_k$. Moreover, the matrices $\mathbf{Y}_k$, $\widehat{\mathbf{H}}_k$ and $\mathbf{T}_k$ are independent of each other, as are the pair $\mathbf{Z}_k$ and $\mathbf{M}_k$. By right multiplying both sides of (30) by $\widehat{\mathbf{F}}_j$, we get

$$\widetilde{\mathbf{H}}_k\widehat{\mathbf{F}}_j = \mathbf{Z}_k\mathbf{M}_k\widehat{\mathbf{F}}_j, \tag{31}$$

for $k \neq j$ due to the fact that $\widehat{\mathbf{H}}_k\widehat{\mathbf{F}}_j = 0$ by the BD procedure. Therefore,

$$\mathbb{E}\left[\widetilde{\mathbf{H}}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\widetilde{\mathbf{H}}_k^{\mathrm{H}}\right] = \mathbb{E}\left[\mathbf{Z}_k\mathbf{M}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\mathbf{M}_k^{\mathrm{H}}\mathbf{Z}_k^{\mathrm{H}}\right]. \tag{32}$$

The inter-user interference in (32) can reach its upper bound in an extreme case, where the channels of all the K users are strongly correlated. In this case, K users share the same clusters around the BS in the ray-based channel model, i.e., $P_1 = P_2 = \cdots = P_K = P$ and $\mathbf{A}_1 = \mathbf{A}_2 = \cdots = \mathbf{A}_K = \mathbf{A}$. Thus, we can omit the subscript k of P_k and $\mathbf{A}_k$ in this proof. As discussed earlier in Sec. IV, the row-spaces of both the feedback channel matrix, $\widehat{\mathbf{H}}_k$, and the spatial direction, $\widetilde{\mathbf{H}}_k$, of user k are distributed in the row-space of $\mathbf{A}$. Since the row-space of $\widetilde{\mathbf{H}}_k$ can be orthogonally decomposed along the row-spaces of $\widehat{\mathbf{H}}_k$ and $\mathbf{M}_k$ as in (30), $\mathbf{M}_k$ should also be distributed in the row-space of $\mathbf{A}$. Hence, by utilizing the asymptotic orthogonality among the row vectors of $\mathbf{A}$ when $M \rightarrow \infty$, $\mathbf{M}_k$ can be expressed as $\mathbf{M}_k = \frac{1}{\sqrt{M}}\mathbf{P}_k\mathbf{A}$, where the rows of $\mathbf{P}_k \in \mathbb{C}^{N\times P}$ are orthonormal. As long as the row-space of the quantized channel matrix $\widehat{\mathbf{H}}_k$ lies in the row-space of $\mathbf{A}$, and according to Sec. III, then the column-space of the precoding matrix $\widehat{\mathbf{F}}_j$ lies in the column-space of $\mathbf{A}^{\mathrm{H}}$. Consequently, utilizing the orthogonality among the column vectors of $\mathbf{A}^{\mathrm{H}}$ when $M \rightarrow \infty$, the precoding matrix $\widehat{\mathbf{F}}_j$ can be expressed as $\widehat{\mathbf{F}}_j = \frac{1}{\sqrt{M}}\mathbf{A}^{\mathrm{H}}\mathbf{E}_j$, where $\mathbf{E}_j \in \mathbb{C}^{P\times m}$ is a unitary matrix whose columns are orthonormal. Hence, substituting in (32), the interference term can be calculated as

$$\begin{aligned} \mathbb{E}\left[\widetilde{\mathbf{H}}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\widetilde{\mathbf{H}}_k^{\mathrm{H}}\right] &= \mathbb{E}\left[\mathbf{Z}_k\mathbf{P}_k\frac{\mathbf{A}\mathbf{A}^{\mathrm{H}}}{M}\mathbf{E}_j\mathbf{E}_j^{\mathrm{H}}\frac{\mathbf{A}\mathbf{A}^{\mathrm{H}}}{M}\mathbf{P}_k^{\mathrm{H}}\mathbf{Z}_k^{\mathrm{H}}\right] \\ &= \mathbb{E}\left[\mathbf{Z}_k\mathbf{P}_k\mathbf{E}_j\mathbf{E}_j^{\mathrm{H}}\mathbf{P}_k^{\mathrm{H}}\mathbf{Z}_k^{\mathrm{H}}\right] \end{aligned} \tag{33}$$

where the second equality follows from the result $\mathbf{A}\mathbf{A}^{\mathrm{H}} = M\mathbf{I}_P$ when $M \rightarrow \infty$.

Lemma 1: In the extreme case that all K users are strongly correlated and share same clusters around the BS, i.e., $P_1 = P_2 = \cdots = P_K = P$ *and* $\mathbf{A}_1 = \mathbf{A}_2 = \cdots = \mathbf{A}_K = \mathbf{A}$, *we have* $\mathbb{E}\left[\mathbf{Z}_k\mathbf{P}_k\mathbf{E}_j\mathbf{E}_j^H\mathbf{P}_k^H\mathbf{Z}_k^H\right] = \frac{N}{P-N}\mathbb{E}\left[\mathbf{Z}_k\mathbf{Z}_k^H\right] = \frac{N}{P-N}\frac{D}{N}$.

Proof: As previously discussed, we have that $\mathbf{M}_k = \frac{1}{\sqrt{M}}\mathbf{P}_k\mathbf{A}$ and, from Sec. IV, the feedback channel matrix can be expressed as $\widehat{\mathbf{H}}_k = \frac{1}{\sqrt{M}}\widehat{\mathbf{G}}_k\mathbf{A}$. Considering that the row-space of $\mathbf{M}_k$ is distributed in the right null space of $\widehat{\mathbf{H}}_k$ as shown in (30), we have

$$\widehat{\mathbf{H}}_k\mathbf{M}_k{}^{\mathrm{H}} = \frac{1}{M}\widehat{\mathbf{G}}_k\mathbf{A}\mathbf{A}^{\mathrm{H}}\mathbf{P}_k{}^{\mathrm{H}} = \widehat{\mathbf{G}}_k\mathbf{P}_k{}^{\mathrm{H}} = 0. \tag{34}$$

Therefore, the row-space of $\mathbf{P}_k$ is distributed in the right null space of $\mathbf{X}_{ik}$. On the other hand, as previously mentioned, the BD precoding matrix can be expressed as $\widehat{\mathbf{F}}_j = \frac{1}{\sqrt{M}}\mathbf{A}^{\mathrm{H}}\mathbf{E}_j$. Since the column-space of the BD precoding matrix $\widehat{\mathbf{F}}_j$ is orthogonal to the row-space of $\widehat{\mathbf{H}}_k$, i.e.,

$$\widehat{\mathbf{H}}_k\widehat{\mathbf{F}}_j = \frac{1}{M}\widehat{\mathbf{G}}_k\mathbf{A}\mathbf{A}^{\mathrm{H}}\mathbf{E}_j = \widehat{\mathbf{G}}_k\mathbf{E}_j = 0. \tag{35}$$

Therefore, the column-space of $\mathbf{E}_j$ is isotropically distributed in the right null space of $\widehat{\mathbf{G}}_k$. Now we have proved that both the row-space of $\mathbf{P}_k$ and the column-space of $\mathbf{E}_j$ are isotropic sub-spaces in the null space of $\widehat{\mathbf{G}}_k$. Based on [39], we have

$$\mathbb{E}\left[\mathbf{Z}_k\mathbf{P}_k\mathbf{E}_j\mathbf{E}_j^{\mathrm{H}}\mathbf{P}_k^{\mathrm{H}}\mathbf{Z}_k^{\mathrm{H}}\right] = \frac{N}{P-N}\mathbb{E}\left[\mathbf{Z}_k\mathbf{Z}_k^{\mathrm{H}}\right] = \frac{N}{P-N}\frac{D}{N}\mathbf{I}_N \tag{36}$$

where D is the average subspace quantization error which, for case I, is given by

$$D = \mathbb{E}\left[d^2(\widetilde{\mathbf{H}}_k, \widehat{\mathbf{H}}_k)\right], \tag{37}$$

while for case II, D is given by

$$D = \mathbb{E}\left[d^2(\mathbf{V}_k, \widehat{\mathbf{V}}_k)\right], \tag{38}$$

where $d(.,.)$ is the chordal distance defined in (10). □

Hence, the interference term is given as

$$\mathbb{E}\left[\widetilde{\mathbf{H}}_k\widehat{\mathbf{F}}_j\widehat{\mathbf{F}}_j^{\mathrm{H}}\widetilde{\mathbf{H}}_k^{\mathrm{H}}\right] = \frac{D}{P-N}\mathbf{I}_N \tag{39}$$

Consequently, the rate gap can be upper bounded using the following equation

$$\Delta R(\gamma) \leq N\log_2\left(1 + \frac{\gamma(K-1)MPN}{Km(P-N)}D\right). \tag{40}$$

B. QUANTIZATION ERROR

In this subsection, we calculate the quantization error, D, of the spatial direction of user k when the AoD-adaptive subspace codebook is used for both cases I & II.

For Case I, we have, from Sec. IV, that $\widehat{\mathbf{H}}_k = \frac{1}{\sqrt{M}}\widehat{\mathbf{G}}_k\mathbf{A}_k$. Also, the spatial direction of the kth user, $\widetilde{\mathbf{H}}_k$, can be represented as $\widetilde{\mathbf{H}}_k = \frac{1}{\sqrt{M}}\widetilde{\mathbf{G}}_k\mathbf{A}_k$ as previously discussed in II-B. Hence, the quantization error, for case I, can be given as

$$D = \mathbb{E}\left[N - \left\|\widetilde{\mathbf{H}}_k\widehat{\mathbf{H}}_k^{\mathrm{H}}\right\|_{\mathrm{F}}^2\right] = \mathbb{E}\left[N - \left\|\frac{1}{M}\widetilde{\mathbf{G}}_k\mathbf{A}_k\mathbf{A}_k^{\mathrm{H}}\widehat{\mathbf{G}}_k^{\mathrm{H}}\right\|_{\mathrm{F}}^2\right] \tag{41}$$

$$\overset{(a)}{\approx} \mathbb{E}\left[N - \left\|\widetilde{\mathbf{G}}_k\widehat{\mathbf{G}}_k^{\mathrm{H}}\right\|_{\mathrm{F}}^2\right]. \tag{42}$$

Step (a) is true due to $\mathbf{A}_k\mathbf{A}_k^{\mathrm{H}} \approx M\mathbf{I}_P$. Both $\widetilde{\mathbf{G}}_k$ and $\widehat{\mathbf{G}}_k$ are isotropically distributed subspaces on the complex P-dimensional space. Then, the distortion or error that results from the quantization of can be bounded as [42]

$$D \leq \bar{D} = \frac{\Gamma(\frac{1}{T})}{T}(C_{PN})^{-\frac{1}{T}}2^{-\frac{B}{T}}, \tag{43}$$

where $T = N(P-N)$ and $C_{PN} = \frac{1}{T!}\prod_{i=1}^{N}\frac{(P-i)!}{(N-i)!}$.

For case II, from Sec. IV, we have that $\mathbf{V}_{k,m} = \frac{1}{\sqrt{M}}\mathbf{A}_k^{\mathrm{H}}\mathbf{J}_k$ and $\widehat{\mathbf{V}}_{k,m} = \frac{1}{\sqrt{M}}\mathbf{A}_k^{\mathrm{H}}\widehat{\mathbf{J}}_k$, where $\mathbf{J}_k \in \mathbb{C}^{P\times m}$ is a unitary matrix. Hence, the quantization error, for case I, can be given as

$$D = \mathbb{E}\left[m - \left\|\mathbf{V}_{k,m}^{\mathrm{H}}\widehat{\mathbf{V}}_{k,m}\right\|_{\mathrm{F}}^2\right] = \mathbb{E}\left[m - \left\|\frac{1}{M}\mathbf{J}_k^{\mathrm{H}}\mathbf{A}_k\mathbf{A}_k^{\mathrm{H}}\widehat{\mathbf{J}}_k\right\|_{\mathrm{F}}^2\right] \tag{44}$$

$$\approx \mathbb{E}\left[m - \left\|\mathbf{J}_k^{\mathrm{H}}\widehat{\mathbf{J}}_k\right\|_{\mathrm{F}}^2\right]. \tag{45}$$

The column-spaces of both $\mathbf{J}_k$ and $\widehat{\mathbf{J}}_k$ are distributed on the P-dimensional space. Then, the quantization error D, for case II, can be bounded as

$$D \leq \bar{D} = \frac{\Gamma(\frac{1}{T})}{T}(C_{Pm})^{-\frac{1}{T}}2^{-\frac{B}{T}}, \tag{46}$$

where $T = m(P-m)$ and $C_{Pm} = \frac{1}{T!}\prod_{i=1}^{m}\frac{(P-i)!}{(m-i)!}$.

C. FEEDBACK BITS

Now, we discuss the required number of feedback bits B that results in a constant rate gap. After bounding the quantization error by $\bar{D}$, the rate loss can be bounded as

$$\Delta R(\gamma) \leq N\log_2\left(1 + \frac{\gamma(K-1)MPN}{Km(P-N)}\bar{D}\right). \tag{47}$$

Let the rate gap be such that $\Delta R(\gamma) \leq \log_2(b)$ bps/Hz, and substituting for $\bar{D}$ from (43), then the number of feedback bits that guarantees this rate loss is given by

$$B = 3.32\,T\log_{10}(\gamma) - T\log_2\left[(b^{\frac{1}{N}} - 1)\frac{Km(P-N)}{(K-1)MPN}\right] + T\log_2\left(\frac{\Gamma(\frac{1}{T})}{T}\right) - \log_2(C), \tag{48}$$

where $C = C_{PN}$ and $T = N(P-N)$, for case I, while $C = C_{Pm}$ and $T = m(P-m)$, for case II. It is noticeable that B scales linearly with the transmit power γ_{dB} in dB.

VI. WATER-FILLING BASED CHANNEL QUANTIZATION AND FEEDBACK

Due to limitations of the zero-forcing techniques in general when the system noise is high, BD based precoding is sub-optimal at low SNR region. Consequently, in this section, we introduce a quantization scheme that is based on optimal power allocation (water-filling) across the multiple data streams to maximize the system's total sum rate at low SNR region. In the case of uniform power allocation, as discussed in the previous sections, only the row-space of $\mathbf{H}_k$, i.e. (the spatial direction of user k) is needed at the BS. However, in order to perform power allocation across data streams, more information about the channel matrix, $\mathbf{H}_k$, is needed at the BS. Therefore, in this section, we present a quantization and feedback scheme of the channel matrix, $\mathbf{H}_k$, which is needed when optimal power allocation (water-filling) across the multiple data streams is intended. The water-filling algorithm needs to know the direction of each channel vector of user k, i.e., the direction of the rows of $\mathbf{H}_k$, as well as the magnitude information of each direction, i.e., the Frobenius norm of each row of $\mathbf{H}_k$. Consequently, we cannot use the subspace quantization codebook that we used in the previous sections, but instead, we shall quantize and feedback each of the N_k channel vectors separately to maintain the direction information in each of them. We will discuss the proposed channel quantization scheme first in the next subsection, then the power allocation scheme is presented in the following subsection. Later, we will show that water-filling is very useful at low SNR regions.

A. PROPOSED CHANNEL QUANTIZER FOR WATER-FILLING

Now, the proposed quantization technique of the channel matrix $\mathbf{H}_k$ of the kth user is discussed. The total number of bits, B, allocated for quantizing $\mathbf{H}_k$ is equally distributed among the rows of $\mathbf{H}_k$. As discussed in Sec. IV, we only need to feedback the path gains matrix $\mathbf{G}_k$ because the channel subspace of each user $\mathbf{A}_k$ is assumed to be known at the BS. Hence, each row of the path gains matrix $\mathbf{G}_k \in \mathbb{C}^{N_k \times P_k}$ is quantized separately using vector quantization. The quantization of the ith row, $\mathbf{g}_{k,i} \in \mathbb{C}^{1\times P_k}$, of the path gains matrix $\mathbf{G}_k$ is chosen from the codebook $\mathcal{C}_{k,i} = \{\mathbf{c}_{k,i,1}, \mathbf{c}_{k,i,2}, \cdots, \mathbf{c}_{k,i,2^{B^*}}\}$, where $\mathbf{c}_{k,i,j} \in \mathbb{C}^{1\times P_k}$ is the jth quantization vector in the codebook and B^* is the number of bits used to quantize each row, $\mathbf{g}_{k,i}$, of $\mathbf{G}_k$. The kth user quantizes its $\mathbf{G}_k$ to N_k quantization vectors that form the quantized path gains matrix $\widehat{\mathbf{G}}_k$, where $\widehat{\mathbf{G}}_k$ is defined as

$$\widehat{\mathbf{G}}_k = \left[\hat{\mathbf{g}}_{k,1}^T, \hat{\mathbf{g}}_{k,2}^T, \cdots, \hat{\mathbf{g}}_{k,N_k}^T\right]^T, \tag{49}$$

and $\hat{\mathbf{g}}_{k,i} = \mathbf{c}_{k,i,Z_{k,i}}$ is chosen from the codebook $\mathcal{C}_{k,i}$, where the index $Z_{k,i}$ is calculated such that

$$Z_{k,i} = \underset{j\in[1,2^{B^*}]}{\arg\min}\, 1 - \left|\mathbf{g}_{k,i}\mathbf{c}_{k,i,j}^{\mathrm{H}}\right|^2. \tag{50}$$

Additionally, the squared Frobenius norm of the ith row of $\mathbf{G}_k$, $\|\mathbf{g}_{k,i}\|_F^2$, is quantized and fed back to the BS. The magnitude information of the channel vectors does contribute in the power allocation solution because it affects the calculation of the singular values of the effective channel, $\mathbf{H}_k\mathbf{F}_k$, of User k, where $\mathbf{F}_k$ is the precoding matrix. We show next, in Sec. VI-B, the procedures of the power allocation solution and how it is computed using the singular values of $\mathbf{H}_k\mathbf{F}_k$. The squared frobenius norm $\|\mathbf{g}_{k,i}\|_F^2$ has an Erlang distriution with parameters $(P_k, 1)$ because the entries of $\mathbf{g}_{k,i} \in \mathbb{C}^{1\times P_k}$ are modeled as complex Gaussian random variables with zero mean and unit variance. Hence, we can quantize $\|\mathbf{g}_{k,i}\|_F^2$, using Lloyd-Max scalar quantizer [43], based on Erlang distribution and number of quantization levels equals to $2^{B_{norm}}$, where B_{norm} is the number of bits allocated for quantization of $\|\mathbf{g}_{k,i}\|_F^2$. Although the frobenius norm information is useful in performing the water-filling process at the BS, we will show in the simulation results section that allocating all the feedback bits in quantizing the direction information of the channel vectors is more useful than allocating a portion of the bits for direction information and another for quantizing the norm information. This reflects that the direction information is more sensitive for enhancing the per-user rate using water-filling than the norm information.

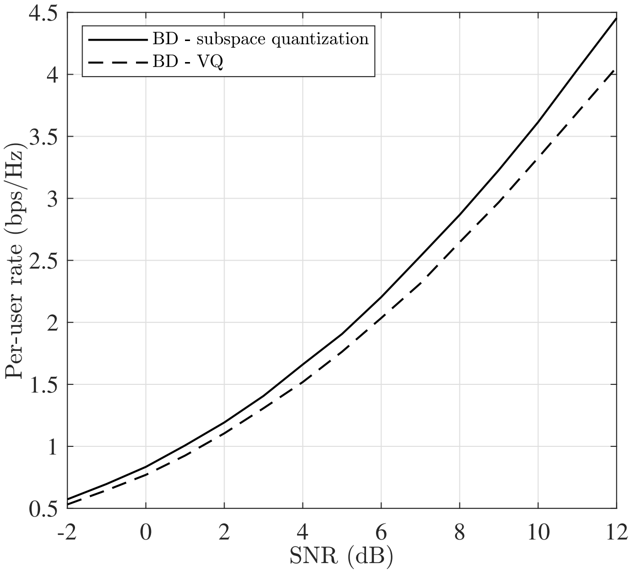


FIGURE 2. Performance comparison between BD with subspace quantization (4 bits/channel subspace) vs. the proposed vector quantization (VQ) without water-filling (2 bits/channel vector), with $N_k = 2$, $P = 3$, $K = 8$, and $M = 128$.

The proposed channel quantizer in this section is suboptimal when uniform power allocation among the users' data streams is adopted. The subspace quantization proposed in Sec. IV is better than this quantizer when the BD with uniform power allocation is used. The reason for this is that the BD procedure does not require the direction information of each row vector of $\mathbf{H}_k$ but only the spatial direction of it as discussed in Sec. II-B i.e. (the subspace spanned by the rows of $\mathbf{H}_k$). Hence, the subspace quantization performs better because it assigns all the bits in quantizing the channel matrix's spatial direction $\widetilde{\mathbf{H}}_k$. However, we show next that the proposed vector quantizer in this section, combined with optimal power allocation across the data streams, outperforms subspace quantization at low-to-mid SNR regions.

Fig. 2 compares the performance of the two different quantizers when uniform power allocation across the data streams is adopted. The figure shows that subspace quantization outperforms the proposed vector quantization in all SNR ranges. This indicates that subspace quantization of $\mathbf{H}_k$ is always better than vector quantization when using BD with uniform power allocation strategy.

B. POWER OPTIMIZATION ALGORITHM

In this subsection, a power optimization technique, which aims to maximize the system's total sum rate at low SNR region, is discussed. Power optimization across the data streams is discussed under both perfect and limited CSI feedback cases.

1) POWER OPTIMIZATION ASSUMING IDEAL CSI AT THE BS

Here, the power optimization algorithm based on ideal CSI at the BS is discussed. We are interested in finding the power allocation diagonal matrices $\Delta_k, \forall k \in \{1, 2, \cdots, K\}$ that maximize the sum rate of the whole system. Hence, the precoding matrix at the BS for user k becomes $\mathbf{F}_k \Delta_k^{1/2}$. When considering the BD based linear precoding at the BS, the inter-user interference is totally cancelled and the sum rate of the system assuming perfect CSI knowledge at the BS becomes

$$R_{\text{tot}} = \mathbb{E}\left\{\sum_{k=1}^{K} \log_2 \left|\mathbf{I}_{N_k} + \mathbf{H}_k \mathbf{F}_k \Delta_k \mathbf{F}_k^{\mathbf{H}} \mathbf{H}_k^{\mathbf{H}}\right|\right\}, \tag{51}$$

where $\mathbf{F}_k$ is the precoding matrix at user k and the diagonal elements of Δ_k scale the power transmitted into each of the columns of $\mathbf{F}_k$. Because of the nature of the BD structure, the BS sees every user as a point-to-point MIMO link. Therefore, the sum information rate of the system can be calculated as

$$R_{\text{tot}} = \mathbb{E}\left\{\sum_{k=1}^{K} \log_2 \left|\mathbf{I}_{N_k} + \Lambda_k^2 \Delta_k\right|\right\}, \tag{52}$$

where Λ_k is a diagonal matrix whose elements $\sigma_{k,i}$ are the singular values of the effective channel, $\mathbf{H}_k \mathbf{F}_k$, of user k and Δ_k is a diagonal matrix whose elements $\delta_{k,i}$ are the power values transmitted into each of the N_k data streams of user k. Now, the power allocation problem that maximizes the sum-rate can be rewritten as

$$\begin{aligned} \max_{\delta_{k,i}} \quad & \sum_{k=1}^{K} \sum_{i=1}^{N_k} \log_2 \left(1 + \sigma_{k,i}^2 \delta_{k,i}\right) \\ \text{s.t.} \quad & \sum_{k=1}^{K} \sum_{i=1}^{N_k} \delta_{k,i} \leq \gamma, \end{aligned}$$

where γ is the total transmitted power at the BS. The above problem is clearly a convex optimization problem and it can be solved using the standard solutions. The solution of this problem is well known and it has a closed form expression [37] when solved using the Lagrange multiplier method. The power allocation solution of the above problem is given as

$$\delta_{k,i}^{*} = \left(\frac{1}{\alpha} - \frac{1}{\sigma_{k,i}^2}\right)^{+}. \tag{53}$$

The value of α is determined such that the total power constraint at the BS is satisfied. Hence, α is the solution of the following equation

$$\sum_{k=1}^{K} \sum_{i=1}^{N_k} \left(\frac{1}{\alpha} - \frac{1}{\sigma_{k,i}^2}\right)^{+} = \gamma. \tag{54}$$

Algorithm 2: Power Optimization Across the Data Streams

```
User $k$ quantizes $\mathbf{G}_k$ into $\widehat{\mathbf{G}}_k$ based on vector
    quantization and send it to the BS
The BS calculates $\widehat{\mathbf{H}}_k = \widehat{\mathbf{G}}_k \mathbf{A}_k$
The BS performs the BD procedures and generates the
    precoding matrices $\widehat{\mathbf{F}}_k$
The BS computes the singular values, $\Lambda_k$, of the
    effective channel, $\widehat{\mathbf{H}}_k \widehat{\mathbf{F}}_k$, of user $k$
The BS calculates the value $\alpha$ that satisfies this equation
    $\sum_{k=1}^{K} \sum_{i=1}^{N_k} \left(\frac{1}{\alpha} - \frac{1}{\hat{\sigma}_{k,i}^2}\right)^{+} = \gamma$
The BS calculates the optimal power scaling values, $\hat{\delta}_{k,i}$,
    according to
    $\hat{\delta}_{k,i}^{*} = \left(\frac{1}{\alpha} - \frac{1}{\hat{\sigma}_{k,i}^2}\right)^{+}$
The final scaled precoding matrix of user $k$ becomes
    $\widehat{\mathbf{F}}_k \Delta_k^{1/2}$
```

2) POWER OPTIMIZATION ALGORITHM CONSIDERING VECTOR QUANTIZATION OF CSI AT THE BS

Now, the power optimization algorithm when considering limited feedback of CSI at the BS is discussed. We shall use the proposed vector quantizer as discussed in Sec. VI-A. The path gains matrix $\mathbf{G}_k$ is first quantized to $\widehat{\mathbf{G}}_k$ at the kth user then fed back to the BS. Then, the BS uses $\widehat{\mathbf{G}}_k$ to form $\widehat{\mathbf{H}}_k$ and hence going through the BD procedure to generate the precoding matrices $\widehat{\mathbf{F}}_k$ as discussed in Sec. III-A1. Then, power optimization accross the data streams is adopted at the BS. Algorithm 2 summarizes these steps in an easy way.

Due to limited feedback of CSI at the BS, the interference on user k due to signals of other users is not totally canceled. Because of this residual interference, we cannot express the information rate of user k as in (52). Hence, the per-user rate considering limited feedback of CSI and power optimization across the data streams is given by

$$\begin{aligned} R_{\text{limited},k}(\Delta_k) = \mathbb{E} \log_2 &\left|\mathbf{I}_{N_k} + \sum_{j=1}^{K} \mathbf{H}_k \widehat{\mathbf{F}}_j \Delta_j \widehat{\mathbf{F}}_j^{\mathbf{H}} \mathbf{H}_k^{\mathbf{H}}\right| \\ &- \mathbb{E} \log_2 \left|\mathbf{I}_{N_k} + \sum_{j=1, j \neq k}^{K} \mathbf{H}_k \widehat{\mathbf{F}}_j \Delta_j \widehat{\mathbf{F}}_j^{\mathbf{H}} \mathbf{H}_k^{\mathbf{H}}\right|. \end{aligned} \tag{55}$$

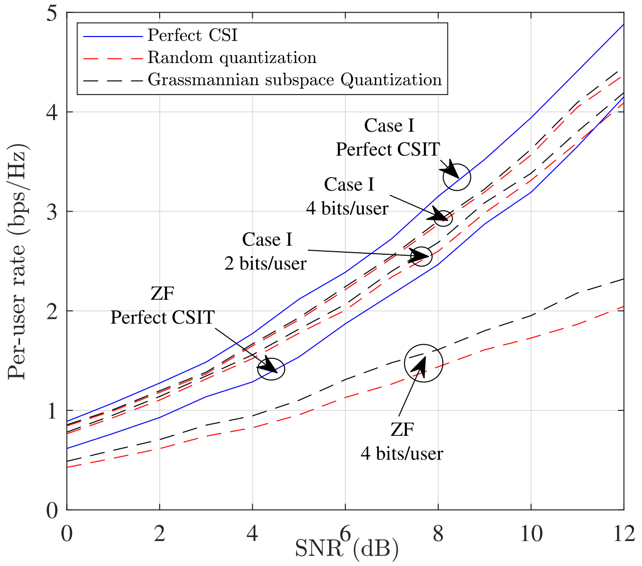


FIGURE 3. BD vs conventional ZF: case I with $N_k = m_k = 2$.

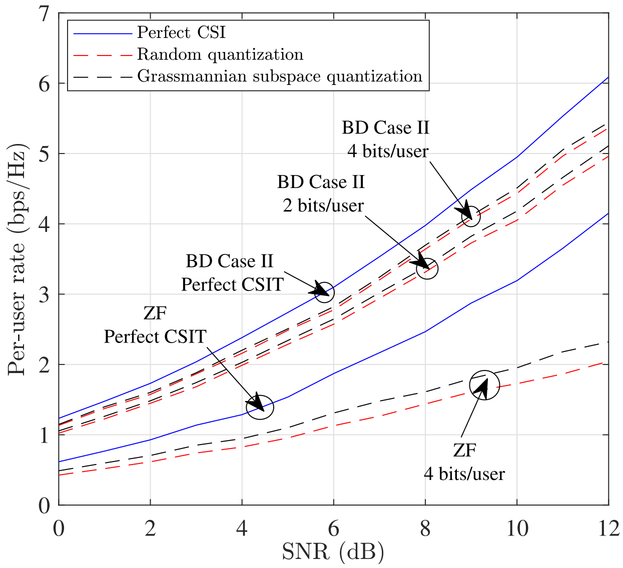


FIGURE 4. BD vs conventional ZF: case II with $N_k = 3$, $m_k = 2$.

TABLE 2. Simulation parameters.

Parameter	Value
BS antennas M	128
Users K	8
Resolvable paths P	3
Case I Rx antennas N_k	2
Case II Rx antennas N_k	3
Data streams to kth user m_k	2
AoDs distribution	$\mathcal{U}\left[-\frac{1}{2}\pi, \frac{1}{2}\pi\right]$

It should be noted that Algorithm (2) uses the quantized channels $\widehat{\mathbf{H}}_k$ and the precoders $\widehat{\mathbf{F}}_k$ to solve a water-filling problem only on the leading term of (55) ignoring the interference term; the interference is taken into the rate expression as a consequence. Therefore, Algorithm (2) can be considered as a simplified approach to solve our water-filling problem without optimizing the whole rate, which is in general a very difficult problem to find a closed-form solution or to optimally solve using a low complexity numerical method.

VII. SIMULATION RESULTS

In this section, the performance of the proposed feedback system and codebook design is examined and verified. The system parameters are set as follows. The number of antennas at the BS is $M = 128$, the number of users in the system is $K = 8$, the number of antennas at each user is $N_k = N = 2$ for case I, while $N = 3$ for case II. The number of data streams transmitted simultaneously to each user is $m_k = 2$ and the number of resolvable paths is $P = 3$. The path AoDs of the users are independent and uniformly distributed in $\left[-\frac{1}{2}\pi, \frac{1}{2}\pi\right]$. The simulation parameters of our simulation setup are collected in Table 2.

A. BD BASED SUBSPACE QUANTIZATION (CASE I & II)

Fig. 3 compares the performance of BD against the conventional ZF scheme, in which the multiple antennas of the same user are orthogonally precoded, in Case I with $N_k = m_k = 2$. Fig. 3 also compares the performance of the ideal case, where perfect CSI is assumed available at the BS, and the limited feedback case where quantized CSI is fed back to the BS with $B = 2$ and 4 per user. Note that in the case of conventional ZF scheme, the channel vector of each antenna at the kth user is separately quantized and fed back to the BS; therefore, the feedback bits for each user are divided among its receive antennas in this case. This is because in the case of conventional ZF, any user antenna is used to receive a single stream, and all other streams must be nulled, including the streams intended for the same user, which is not the case in BD. In Fig. 3, we plot the per-user rate using the AoD-adaptive codebook with both random subspace quantization and using Grassmannian subspace packing based codebook. From this figure, we can easily see the BD approach's performance gains compared to the conventional ZF approach. It can also be noticed that Grassmannian codes are always better (or slightly better) than random codes. Finally, it is clear that increasing the number of feedback bits enhances the system performance, and we can get arbitrary close to the performance of the ideal system with perfect CSI at the BS.

Remark: It should be noted that the random subspace quantization scheme does not correspond to a fixed quantizer that is picked randomly and used in all iterations. Random quantization here means that in each iteration, we generate a quantizer whose quantization subspaces are drawn randomly from a uniform distribution over the space. Then, the per-user rate is averaged over all the iterations assuming that the quantizer is known at both the user and BS side. By averaging over a large number of iterations, some good quantizers might equalize the bad performance of some other bad quantizers. Indeed, random quantization is impractical to be

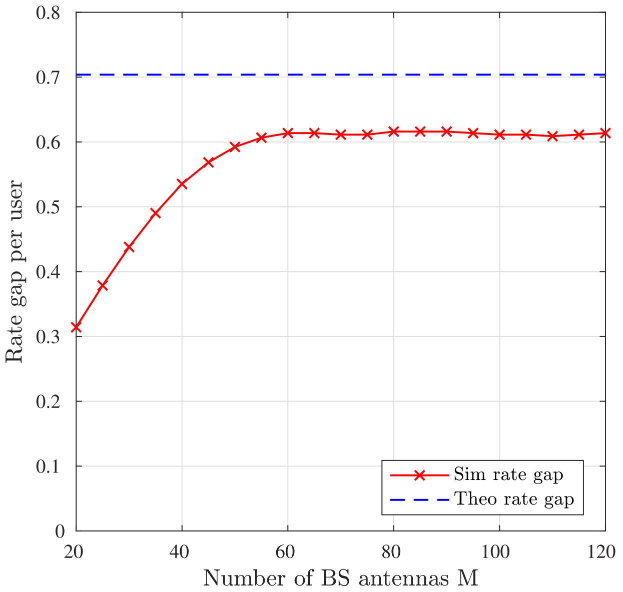


FIGURE 5. Rate gap analytical upper bound in (47) vs simulated rate gap as the number of BS antennas M increase, downlink SNR = 10 dB.

implemented in any practical system and it is only here used to evaluate the feedback scheme and for comparison purpose. In contrast, Grassmannian codes are structured, which deem them suitable for practical implementation in real systems as they are designed according to some specific design criteria with some performance guarantees.

Fig. 4 compares the performance of BD with ideal and quantized CSI against the ideal and quantized CSI of the conventional ZF scheme for case II with $N_k = 3$, $m_k = 2$ with $B = 2$ and 4 per user. The same observations mentioned above while commenting on the results of Fig. Fig. 3 apply in this case as well. Moreover, it is noticeable that case II has a higher per-user rate than case I for the same number of user streams. This is since in case II, we assume more receiving antennas at each user than the number of streams, which introduces diversity gain at the users.

Then, we assess the derived analytical rate gap upper bound in (47) versus the number of BS antennas M. The number of feedback bits B is set as (48) with $b = 3$ and the downlink SNR at the users is fixed at 10 dB. Fig. 5 shows the theoretically derived upper bound on rate gap in (47) and the simulated rate gap of the proposed feedback scheme, which is calculated by $\Delta R = R_{CSIT} - R_{QUANT}$. The graph shows that the analytical rate gap becomes tighter as the number of BS antennas increases. This observation makes sense because the assumption that $\mathbf{A}_k\mathbf{A}_k^{\mathrm{H}} \approx M\mathbf{I}_{P_k}$, which we used to derive the theoretical upper bound, becomes more valid.

In Fig. 6 and Fig. 7, we present numerical results for the practical per-user rate when using a random quantization codebook for both cases I and case II respectively. The required number of feedback bits is scaled as in (48) to guarantee a maximum rate gap of $\log_2(b)$, where we show the results for $b = 2$ and 4. We notice in Fig. 6 and Fig. 7 that the rate gap, between the ideal case (perfect CSI at the BS) and the practical case, does not increase as the SNR increases; this is due to scaling the number of feedback bits B with the transmitted power γ_{dB} as explained above. It is clear that the rate gap at any SNR does not exceed the maximum value of $\log_2(b)$, which validates the expression in (48) for both cases I & II.

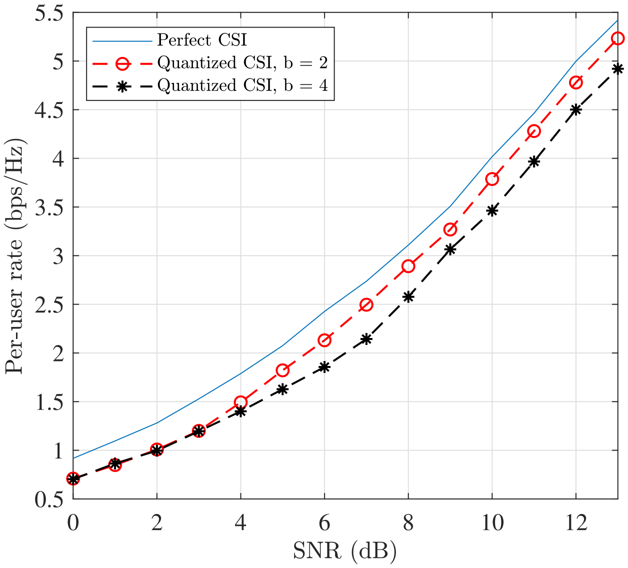


FIGURE 6. Ideal vs quantized CSI, case I, with scaled B as in (48).

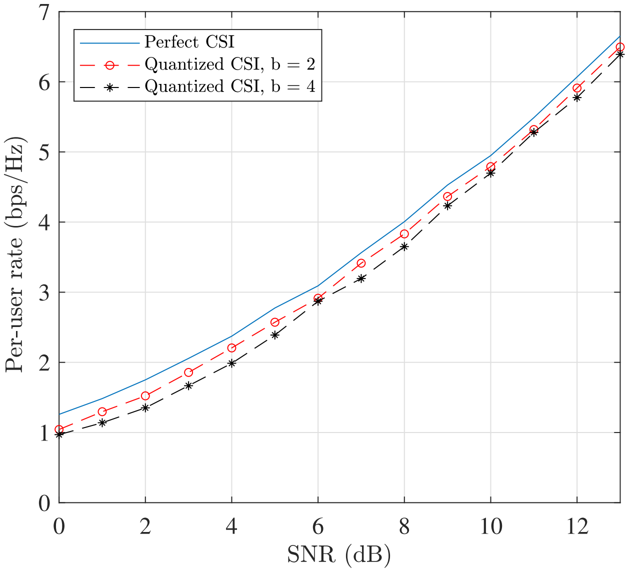


FIGURE 7. Ideal vs quantized CSI, case II, with scaled B as in (48).

B. BD WITH WATER-FILLING BASED VECTOR QUANTIZATION

In this subsection, the derived power allocation scheme's performance based on the proposed vector quantizer is examined and verified. Also, we compare the power allocation scheme based on vector quantization with the subspace quantization

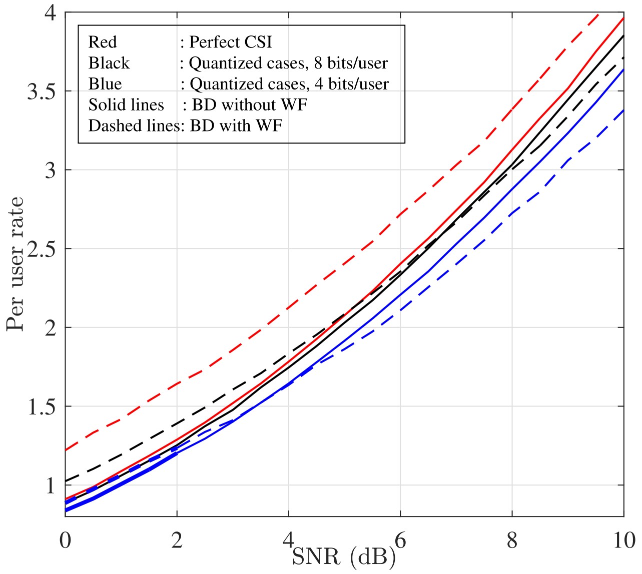


FIGURE 8. Performance of BD with and without WF for both ideal and quantized cases with $B = 4, 8$ bits/user.

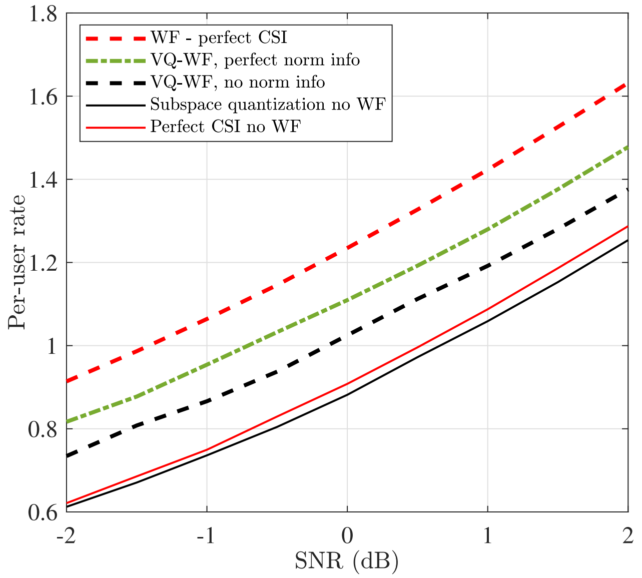


FIGURE 9. Performance of BD-subspace quantization vs. BD-VQ combined with WF at low SNR region with 8 bits/user.

technique discussed in Sec. IV. The number of antennas at each user, $N_k = N = 2$, is equal to the number of data streams transmitted simultaneously to each user, $m_k = 2$.

In Fig. 8, the performance of the BD scheme without power optimization is compared to the performance of BD with power optimization for both ideal and quantization cases. We use a total of $B = 8$ and 4 bits per user, whether for subspace quantization or vector quantization. The graph clearly shows that ideal BD with power optimization (water-filling) outperforms the regular ideal BD for all SNR ranges. However, the gap between both schemes decreases in higher SNR regions. The reason for this fact is that water-filling resists the system noise. Hence, at high SNR values, the system noise nearly becomes insignificant, and the regular BD approaches BD with water-filling.

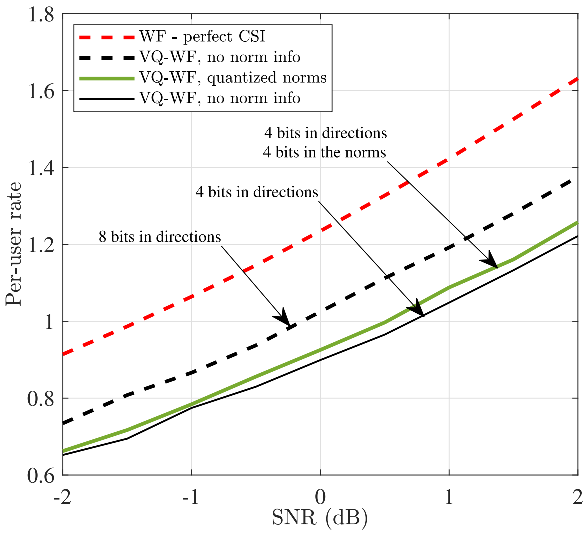


FIGURE 10. Performance comparison of BD-VQ combined with WF when only quantizing the direction information vs. quantizing both directions and norms.

Fig. 8 also shows a trade-off between both schemes in the case of limited feedback (quantization). In the low SNR regime, water-filling with the proposed vector quantization scheme is better than the regular BD with subspace quantization. The reason for this trade-off is that we have two sources of noise, in that case, the system noise and the quantization noise. At low SNR, the system noise is more dominant than the quantization noise, hence using water-filling along with vector quantization is more useful as it is more immune to the system noise. However, at high SNR, the noise resulted from the quantization of $\mathbf{H}_k$ is more dominant. Hence, no gain is added from using the water-filling solution, and subspace quantization, which is the better quantizer when no water-filling as in Fig. 2, resulted in better performance.

As long as BD based power optimization scheme is useful at the low SNR regime, Fig. 9 shows a comparison of both schemes in the range from SNR $= -2dB$ to SNR $= 2dB$ with $B = 8$ bits/user. The graph shows that BD based water-filling using the proposed vector quantization clearly outperforms the regular BD based subspace quantization at these low SNR levels. There is nearly a gain of 1dB in SNR when using water-filling with vector quantization. This gain is moderately good at these low SNR levels. Fig. 9 also shows the effect of having the ideal Frobenius norms of the channel vectors $\mathbf{g}_{k,i}$ on the water-filling solution at the BS. It is clear that this norm information does enhance the BD based vector quantization with water-filling because it affects the calculation of the singular values of the effective channels of the users as discussed before. However, the performance of the water-filling scheme is still better than the regular BD based subspace quantization if all the bits are put in quantizing only the directions of $\mathbf{g}_{k,i}$ and no information is fed back about the norms of $\mathbf{g}_{k,i}$.

Finally, we show in Fig. 10 that there is no gain of quantizing the Frobenius norms of $\mathbf{g}_{k,i}$ when we have a limited number of feedback bits per user, but we shall put all the bits in quantizing only the directions. As previously discussed in Sec. VI-A, we used the Lloyd-Max algorithm to quantize $\|\mathbf{g}_{k,i}\|_F^2$ which has an Erlang distribution. The performance of BD based vector quantization with water-filling is better when we allocate the total $B = 8$ bits for quantizing the directions of $\mathbf{g}_{k,i}$ than allocating 4 bits for the directions and another 4 bits for the squared norm values.

VIII. CONCLUSION

In this article, we have considered the problem of channel feedback in FDD massive MIMO systems with multiple antennas at the users. We devised a channel feedback scheme using low-dimensional codebooks to reduce the required feedback bits. The rate gap between quantized channel feedback and the case with perfect CSI at the BS has been mathematically quantified for the considered cases. A systematic approach to design the channel feedback codebooks, in which the codebook design is formulated as a subspace packing problem over the Grassmannian manifold, was proposed. It was shown that using vector quantization, combined with water-filling, can outperform BD-based subspace quantization in the low SNR region. However, in the high SNR region, the situation is reversed, and the performance of BD-based subspace quantization becomes better where the effect of receiver noise is less significant, while the effect of the quantization noise becomes dominant.

For future directions of this research, we may consider studying and analyzing the use of analog feedback for the path gain matrix $\mathbf{G}_k$ and compare it to the *digital* subspace quantization proposed in this paper. It is expected to have some trade-off between both schemes in the high/low SNR regions that needs to be investigated.

REFERENCES

[1] 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.

[2] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Jun. 2014.

[3] T. L. Marzetta and B. M. Hochwald, "Fast transfer of channel state information in wireless systems," *IEEE Trans. Signal Process.*, vol. 54, no. 4, pp. 1268–1278, Apr. 2006.

[4] Qualcomm. (2013). *FDD/TDD Comparison Key Messages*. [Online]. Available: https://www.qualcomm.com/documents/fddtdd-comparison

[5] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169–6183, Dec. 2015.

[6] Z. Lv and Y. Li, "A channel state information feedback algorithm for massive MIMO systems," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1461–1464, Jul. 2016.

[7] X. Rao and V. K. N. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, Jun. 2014.

[8] A. Liu, F. Zhu, and V. K. N. Lau, "Closed-loop autonomous pilot and compressive CSIT feedback resource adaptation in multi-user FDD massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 1, pp. 173–183, Jan. 2017.

[9] M. B. Khalilsarai, S. Haghighatshoar, X. Yi, and G. Caire, "FDD massive MIMO via UL/DL channel covariance extrapolation and active channel sparsification," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 121–135, Jan. 2019.

[10] P. Liang, J. Fan, W. Shen, Z. Qin, and G. Y. Li, "Deep learning and compressive sensing-based CSI feedback in FDD massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 9217–9222, Aug. 2020.

[11] T. Wang, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based CSI feedback approach for time-varying massive MIMO channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 416–419, Apr. 2019.

[12] M. S. Sim, J. Park, C.-B. Chae, and R. W. Heath, Jr., "Compressed channel feedback for correlated massive MIMO systems," *J. Commun. Netw.*, vol. 18, no. 1, pp. 95–104, Feb. 2016.

[13] W. Shen, L. Dai, Y. Li, Z. Wang, and L. Hanzo, "Channel feedback codebook design for millimeter-wave massive MIMO systems relying on lens antenna array," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 736–739, Oct. 2018.

[14] S. Schwarz, M. Rupp, and S. Wesemann, "Grassmannian product codebooks for limited feedback massive MIMO with two-tier precoding," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 1119–1135, Sep. 2019.

[15] F. Rottenberg, T. Choi, P. Luo, C. J. Zhang, and A. F. Molisch, "Performance analysis of channel extrapolation in FDD massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2728–2741, Jan. 2020.

[16] W. Shen, L. Dai, B. Shim, Z. Wang, and R. W. Heath, Jr., "Channel feedback based on AoD-adaptive subspace codebook in FDD massive MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5235–5248, Jun. 2018.

[17] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[18] M. A. AlaaEldin, K. G. Seddik, and W. Mesbah, "AoD-adaptive channel feedback in FDD massive MIMO systems with multiple-antenna users," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 206–212.

[19] M. AlaaEldin, "Channel feedback in FDD massive MIMO systems with multiple-antenna users," M.S. thesis, AUC Knowl. Fountain, Amer. Univ. Cairo, New Cairo, Egypt, 2019.

[20] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[21] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[22] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, Jun. 2014.

[23] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[24] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[25] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836–869, 2nd Quart., 2018.

[26] T. Bai, A. Alkhateeb, and R. W. Heath, Jr., "Coverage and capacity of millimeter-wave cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 70–77, Sep. 2014.

[27] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[28] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2013.

[29] K. Venugopal, A. Alkhateeb, N. González-Prelcic, and R. W. Heath, Jr., "Channel estimation for hybrid architecture-based wideband millimeter wave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1996–2009, Sep. 2017.

[30] V. Raghavan and A. M. Sayeed, "Sublinear capacity scaling laws for sparse MIMO channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 345–364, Jan. 2011.

[31] A. M. Sayeed and V. Raghavan, "Maximizing MIMO capacity in sparse multipath with reconfigurable antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 1, pp. 156–166, Jun. 2007.

[32] H. Xie, F. Gao, and S. Jin, "An overview of low-rank channel estimation for massive MIMO systems," *IEEE Access*, vol. 4, pp. 7313–7321, 2016.

[33] P. Zhao, Z. Wang, and C. Sun, "Angular domain pilot design and channel estimation for FDD massive MIMO networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[34] J. Choi, D. J. Love, and P. Bidigare, "Downlink training techniques for FDD massive MIMO systems: Open-loop and closed-loop training with memory," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 802–814, Oct. 2014.

[35] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

[36] I. S. Dhillon, R. W. Heath, Jr., T. Strohmer, and J. A. Tropp, "Constructing packings in Grassmannian manifolds via alternating projection," *Exp. Math.*, vol. 17, no. 1, pp. 9–35, 2008.

[37] G. C. Raleigh and J. M. Cioffi, "Spatio-temporal coding for wireless communication," *IEEE Trans. Commun.*, vol. 46, no. 3, pp. 357–366, Mar. 1998.

[38] D. J. Love and R. W. Heath, Jr., "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2967–2976, Aug. 2005.

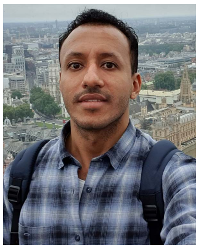
[39] N. Ravindran and N. Jindal, "Limited feedback-based block diagonalization for the MIMO broadcast channel," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1473–1482, Oct. 2008.

[40] V. Va, J. Choi, and R. W. Heath, Jr., "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5014–5029, Jun. 2017.

[41] W. Santipach and M. L. Honig, "Asymptotic capacity of beamforming with limited feedback," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Jun. 2004, p. 290.

[42] W. Dai, Y. Liu, and B. Rider, "Quantization bounds on Grassmann manifolds and applications to MIMO communications," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1108–1123, Mar. 2008.

[43] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inf. Theory*, vol. 6, no. 1, pp. 7–12, Mar. 1960.



MAHMOUD ALAAELDIN (Member, IEEE) received the B.Sc. degree in electrical engineering from Alexandria University, Alexandria, Egypt, in 2013, and the M.Sc. degree from The American University in Cairo (AUC), Cairo, Egypt, in 2019. He is currently pursuing the Ph.D. degree with the Electrical and Electronic Engineering Department, The University of Manchester, Manchester, U.K. Through his years at AUC, he was a Graduate Fellow and a Teacher Assistant with the Electronics and Communications Department. He is currently a Researcher with the Electrical and Electronic Engineering Department, The University of Manchester. He received a University Fellowship from AUC, in 2017 and 2018. In 2019, he received a Research Fellowship with The University of Manchester under the Marie Sklodowska-Curie Actions grant scheme, which is funded by the European Union's Horizon 2020 Research and Innovation Program. His research interests include wireless communications, signal processing, multiuser MIMO/OFDM systems, massive MIMO systems, channel estimation and feedback, physical layer network coding, intelligent reflecting surfaces, optimization, and applications of machine learning in wireless communications.

EMAD ALSUSA (Senior Member, IEEE) received the Ph.D. degree in telecommunications from the University of Bath, U.K., in 2000. In 2000, he was appointed to work on developing high data rates systems as part of an industrial project based at Edinburgh University. In September 2003, he joined The Manchester University (then UMIST) as a Faculty Member, where his current rank is a Reader with the Communication Engineering Group. His research interests include communication systems with a focus on physical, MAC and network layers, including developing techniques and algorithms for array signal detection, channel estimation and equalization, adaptive signal precoding, interference avoidance through novel radio resource management techniques, cognitive, energy and spectrum optimization techniques, cellular networks, the IoT, industry 4.0, and powerline communications. His research work has resulted in over 200 journals and refereed conference publications mainly in top IEEE Transactions and Conferences. He has supervised over 30 Ph.D.'s to successful completion. He is also an Editor of the IEEE Wireless Communication Letters, a fellow of the U.K. Higher Academy of Education, and the TPC Track Chair of a number of conferences, such as VTC'16, GISN'16, PIMRC'17, and Globecom'18, and the General Co-Chair of the OnlineGreenCom'16 Conference. He is currently the U.K. representative with the International Union of Radio Science, and the Co-Chair of the IEEE Special Working Group on RF Energy Harvesting. He has received a number of awards, including the Best Paper Award in the International Symposium on Power Line Communications 2016 and the Wireless Communications and Networks Conference 2019.

KARIM G. SEDDIK (Senior Member, IEEE) received the B.Sc. (Hons.) and M.Sc. degrees in electrical engineering from Alexandria University, Alexandria, Egypt, in 2001 and 2004, respectively, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2008. He is currently a Professor with the Electronics and Communications Engineering Department, The American University in Cairo (AUC), and the Associate Dean for Graduate Studies and Research, School of Sciences and Engineering (SSE), AUC. Before joining AUC, he was an Assistant Professor with Alexandria University. His research interests include applications of machine learning in communication networks, intelligent reflecting surfaces, age of information, cognitive radio communications, and layered channel coding. He has served on the technical program committees of numerous IEEE conferences in the areas of wireless networks and mobile computing.

He was a recipient of The American University in Cairo Faculty Merit Award for Excellence in Research and Creative Endeavors, in 2021. He is a recipient of the State Encouragement Award, in 2016, and the State Medal of Excellence, in 2017. He is also a recipient of the Certificate of honor from the Egyptian President for being ranked first among all departments with the College of Engineering, Alexandria University, in 2002. He received the Graduate School Fellowship from the University of Maryland, in 2004 and 2005, and the Future Faculty Program Fellowship from the University of Maryland, in 2007. He also coauthored a conference paper that received the best conference paper award from the IEEE Communication Society Technical Committee on Green Communications and Computing, in 2019.

WESSAM MESBAH (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees (Hons.) in electrical engineering from Alexandria University, Alexandria, Egypt, in 2000 and 2003, respectively, and the Ph.D. degree from McMaster University, Hamilton, ON, Canada, in 2008. From 2009 to 2010, he was a Postdoctoral Research Fellow with Texas A&M University, Doha, Qatar. He joined the Electrical Engineering Department, King Fahd University of Petroleum and Minerals, in 2010, where he is currently an Associate Professor. His research interests include cooperative communications and relay channels, layered multimedia transmission, wireless sensor networks, multiuser MIMO/OFDM systems, cognitive radio, optimization, game theory, smart metering, and smart grids.