Feedback Mechanisms for FDD Massive MIMO with D2D-based Limited CSI Sharing

Junting Chen, *Member, IEEE*, Haifan Yin,
Laura Cottatellucci, *Member, IEEE*, David Gesbert, *Fellow, IEEE*

***Abstract*—Channel state information (CSI) feedback is a challenging issue in frequency division duplexing (FDD) massive MIMO systems. This paper studies a cooperative feedback scheme, where the users first exchange their CSI with each other through device-to-device (D2D) communications, then compute the precoder by themselves, and feed back the precoder to the base station (BS). Analytical results are derived to show that the cooperative precoder feedback is more efficient than the CSI feedback in terms of interference mitigation. To reduce the delays for CSI exchange, we develop an adaptive CSI exchange strategy based on signal subspace projection and optimal bit partition. Numerical results demonstrate that the proposed cooperative precoder feedback scheme with adaptive CSI exchange significantly outperforms the CSI feedback scheme, even under moderate delays for CSI exchange via D2D.**

***Index Terms*—Massive MIMO, device-to-device, limited feedback, precoder feedback, subspace projection**

I. Introduction

Massive multiple-input multiple-output (MIMO) is widely considered to be one of the key enabling technologies for future wireless communication systems [1]–[4]. With more antennas at the BS, massive MIMO systems have more degrees of freedom to exploit for spatial multiplexing and interference suppression. However, realizing such performance gain requires additional efforts on acquiring the CSI, which has a large dimension. A number of works focus on time-division duplex (TDD) systems, where channel reciprocity can be exploited to obtain the downlink CSI from the uplink pilots transmitted by the users [5]–[7]. However, FDD systems are still dominant in current cellular networks [8], [9].

Conventional limited feedback schemes in correlated channels rely on pre-defined codebooks to quantize and feedback the channel vector [10]–[14]. However, these methods are not scalable to massive MIMO, because the size of the codebook is exponential to the number of feedback bits, which should increase linearly with the number of transmit antennas in order to realize the full multiplexing gain [15]. Hence designing improved feedback schemes in the context of FDD massive MIMO is both challenging and timely. Among the state-of-the-art feedback schemes, trellis-coded quantizers were studied in [16], [17] for massive MIMO with moderate to high feedback loading, using source coding techniques with only a small codebook. Compressive sensing techniques were applied to channel estimation and feedback in [18], [19], under the sparsity assumption for the massive MIMO channel. In contrast to developing vector quantization and reconstruction techniques for massive MIMO, rate splitting encoding strategies at the massive MIMO transmitter were designed in [20], [21] to reinforce robustness through encoding under limited CSI feedback. Furthermore, a two-layer precoding structure was introduced in [22]–[25] to relieve the burden of instantaneous CSI feedback by exploiting the low rank property of the channel covariance matrices. However, the low rank property may not exist in some propagation scenarios due to possible rich scattering environment and sufficient antenna spacing.

In this paper, we tackle the CSI limited feedback issue in FDD massive MIMO systems by exploiting user-level cooperation among nearby users,[1] where users first exchange the (quantized) instantaneous CSI with each other via D2D communications, compute the precoder based on the imperfect global CSI, and then feedback the precoder (rather than the CSI) to the BS. Note that with the rapid development of D2D communications, it is feasible to exchange information with the nearby users in low power[2] and low latency. In particular, the power consumption to exchange a limited amount of CSI over D2D could be negligible due to the small path loss to the nearby users as compared to communicating to the BS that is typically far away.

Similar structure that exploits user cooperation via D2D was considered in [26]–[28], where users share data for receive decoding. In [29], users share the CSI for feedback reduction in user scheduling. By contrast, we propose CSI (instead of data) sharing for precoder feedback that could substantially increase the throughput in massive MIMO downlink. The intuition of the precoding feedback is that, first, experience and analysis have shown that feedback resources for MIMO precoding are better used to convey information directly in the

Manuscript received November 21, 2016; revised March 2017; accepted May 4, 2017.

This research was supported by Huawei France Research Center. The material in this paper was presented, in part, at the Asilomar Conference on Signals, Systems, and Computers, Asilomar, CA, USA, November 2015.

Junting Chen was with Department of Communication Systems, EURECOM, Sophia-Antipolis, France. He is now with Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA (e-mail:juntingc@usc.edu).

Haifan Yin was with Department of Communication Systems, EURECOM, Sophia-Antipolis, France. He is now with Sequans Communications, Paris, France.

Laura Cottatellucci and David Gesbert are with Department of Communication Systems, EURECOM, Sophia-Antipolis, France (e-mail: {cottatel, gesbert}@eurecom.fr).

[1] For the users that are far apart, they can be separated into groups and treated individually using two-layer precoding techniques [22], [23].

[2] For example, when the users are 10 meters to each other and 60 meters to the BS, the D2D link has path loss 20 dB less than the feedback link to the BS.

precoder domain rather than in the channel domain. This is because greater mismatch may be brought in by computing the MIMO precoder from the quantized CSI feedback as quantization errors propagate during channel inversion [30]. Second, computing the precoder at the user side was not possible in classical MIMO systems without D2D, but it is feasible when D2D is exploited. In the ideal case of perfect D2D, CSI exchange allows the users to obtain the global CSI, compute, and feed back the precoder to the BS. Significant throughput gain of such precoder feedback scheme has been demonstrated in prior work [31].

However, CSI exchange over D2D induces delays, and hence it is not realistic to exchange the unquantized CSI among users as studied in the prior work [31]. From the Little's law that the average delay is proportional to the packet size, it is important to limit the amount of bits for CSI exchange over D2D. Therefore, we are interested in the following two fundamental questions: (i) *Does the precoder feedback scheme work under imperfect CSI exchange,* and (ii) *How to efficiently quantize and exchange the CSI via D2D?* Some preliminary analytical results under limited D2D channel capacity were presented in [32], but it is still not known how to efficiently exchange the CSI among the users for the limited rate precoder feedback.

To address these challenging issues, we develop strategies and analytical results for two application scenarios of the cooperative precoder feedback scheme. In the first scenario, we consider the users have uncorrelated channels with identical path loss, and we analyze the performance under limited CSI exchange. In the second scenario, we consider that the users have non-identical channel statistics, where the users may experience different path loss or have different signal subspaces. We propose a novel efficient CSI exchange strategy and derive the optimal bit partition over each D2D link to achieve the minimum interference leakage for the proposed cooperative precoder feedback scheme. The key intuition is that, the users only need to share the portion of CSI that lies in the overlapping signal subspace. For example, in the extreme case when two users have non-overlapping signal subspaces, they do not need to exchange the CSI. In the other extreme case when two users have identical signal subspace and identical path loss, they need high quality CSI exchange. With this, a significant amount of bits can be saved from transmitting over the D2D, which may result in a substantial decrease of the delay for CSI exchange.

The major findings and contributions of this paper are summarized as follows:

- We develop a user-level cooperative precoder feedback framework based on CSI exchange among users, and we propose a novel CSI exchange strategy, which is demonstrated to save one third of the bits for CSI exchange and to be robust under up to moderate CSI exchange delays.
- We analyze the performance of the cooperative precoder feedback scheme under limited but sufficient number of bits for CSI exchange via D2D. We found that the proposed scheme can reduce the interference leakage to $1/(K-1)$ of the CSI feedback scheme in a K-user system under uncorrelated MIMO channels with identical path loss.
- We demonstrate that even with limited D2D capacity and moderate delays for CSI exchange, the proposed scheme can significantly outperform the CSI feedback scheme, where it saves up to half of the bits for the feedback to the BS for the same throughput performance.

Table I
SUMMARY OF IMPORTANT NOTATIONS

Symbols	Meanings
B_{f}	Bits per user to feedback to the BS
b_{kj}	Bits for user k to transmit the CSI to user j
B_{c}	We treat $b_{kj} = B_{\text{c}}$ in Section III for analytical results
B_{tot}	Total number of bits for CSI exchange, $B_{\text{tot}} = \sum_{k,j} b_{kj}$
$\hat{\mathbf{h}}_k$	The quantized channel of $\mathbf{h}_k$
$\hat{\mathbf{g}}_k$	The quantized channel direction of $\mathbf{h}_k/\|\mathbf{h}_k\|$
$\mathbf{h}_k^{(j)}$	The portion of channel $\mathbf{h}_k$ projected onto user j's subspace
$\mathbf{g}_k^{(j)}$	The normalized $\mathbf{h}_k^{(j)}$ by the path loss (eq. (18))
$\mathbf{U}_k$	Matrix that contains the dominant eigenvectors of $\mathbf{R}_k$
$\mathbf{U}_{kj}$	Matrix that contains the dominant eigenvectors of $\mathbf{U}_j^{\text{H}}\mathbf{R}_k\mathbf{U}_j$

The rest of the paper is organized as follows. Section II introduces the precoder feedback scheme with the CSI exchange mechanism. Section III analyzes the interference leakage under uncorrelated channels with identical path loss. Section IV studies the efficient CSI exchange strategy under non-identical channels, where users experience different signal subspaces and path loss. Numerical results are demonstrated in Section V and conclusions are given in Section VI.

Notations: The notations $\|\mathbf{a}\|$ and $\|\mathbf{A}\|$ denote the Euclidean norm of vector $\mathbf{a}$ and the matrix 2-norm of $\mathbf{A}$, respectively. In addition, $(\cdot)^{\text{H}}$ denotes the Hermitian transpose and $\text{tr}\{\mathbf{A}\}$ denotes the trace of matrix $\mathbf{A}$. Furthermore, important symbols are summarized in Table I.

II. System Model

In this section, we elaborate the system model for massive MIMO downlink transmission and introduce the cooperative precoder feedback based on user-level cooperation.

A. Signal Model

Consider a single cell massive MIMO network, where the BS equips with N_t antennas and serves K users. Denote the downlink channel of user k as $\mathbf{h}_k^{\text{H}}$, where $\mathbf{h}_k \in \mathbb{C}^{N_t}$ is a column vector and is independent across users. The received signal of user k is given by

$$y_k = \sqrt{\frac{P}{K}}\mathbf{h}_k^{\text{H}}\mathbf{w}_k s_k + \sqrt{\frac{P}{K}}\sum_{j\neq k}\mathbf{h}_k^{\text{H}}\mathbf{w}_j s_j + n_k$$

where s_k is the transmitted symbol with $\mathbb{E}\{|s_k|^2\} = 1$, $\mathbf{w}_k \in \mathbb{C}^{N_t}$ is the precoder with $\|\mathbf{w}_k\| = 1$, $n_k \sim \mathcal{CN}(0, 1)$ is the additive Gaussian noise, and P is the total transmission power.

Assume that $\mathbf{h}_k$ follows distribution $\mathcal{CN}(\mathbf{0}, l_k\mathbf{R}_k)$, where the covariance matrix $\mathbf{R}_k$ is normalized to $\text{tr}\{\mathbf{R}_k\} = N_t$ and

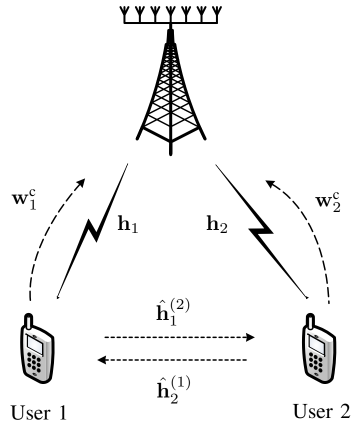

Figure 1. A signaling example of the cooperative precoder feedback scheme in two-user case.

l_k denotes the path loss. The statistics $\{l_k, \mathbf{R}_k\}$ is assumed to be static and known by all the users. Perfect CSI $\mathbf{h}_k$ is assumed available at each user k. In addition, consider that the users exploit reliable D2D communication links for finite rate CSI exchange with each other. The CSI exchange and the feedback strategies are specified as follows.

B. Cooperative Precoder Feedback based on CSI Exchange

Consider the system is operated in FDD mode and explicit feedback is required for CSI acquisition and downlink precoding. Suppose each user has $B_{\rm f}$ bits for the feedback to the BS. In conventional CSI feedback, each user quantizes the channel $\mathbf{h}_k$ into $\hat{\mathbf{h}}_k$ coded by $B_{\rm f}$ bits and feeds back $\hat{\mathbf{h}}_k$ to the BS. Based on the global CSI $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \dots, \hat{\mathbf{h}}_K]$, the BS computes the precoding matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_K]$.

In this paper, we consider a *cooperative feedback* scheme, which consists of two phases:

- *CSI Exchange:* Each user k employs a quantizer $\mathcal{Q}_{kj}$ to share the quantized channel using b_{kj} bits
$$\hat{\mathbf{h}}_k^{(j)} = \mathcal{Q}_{kj}(\mathbf{h}_k)$$
to user j via D2D communication. Where the quantization error $\mathbf{h}_k - \hat{\mathbf{h}}_k^{(j)}$ is assumed to be zero mean and uncorrelated to $\hat{\mathbf{h}}_k^{(j)}$. After the CSI exchange, each user k knows the imperfect global CSI
$$\hat{\mathbf{H}}_k = [\hat{\mathbf{h}}_1^{(k)}, \hat{\mathbf{h}}_2^{(k)}, \dots \hat{\mathbf{h}}_{k-1}^{(k)}, \mathbf{h}_k, \hat{\mathbf{h}}_{k+1}^{(k)}, \dots, \hat{\mathbf{h}}_K^{(k)}]. \quad (1)$$
- *Cooperative Feedback:* With the global CSI, each user first computes the precoder
$$\mathbf{w}_k^{\rm c} = \mathcal{W}_k(\hat{\mathbf{H}}_k)$$
and then feeds back the precoder $\mathbf{w}_k^{\rm c}$ to the BS using $B_{\rm f}$ bits.

The BS applies the precoding vectors $\mathbf{w}_k = \mathbf{w}_k^{\rm c}$ for downlink transmission. Fig. 1 illustrates a signaling example of the cooperative precoder feedback scheme in two-user case.

Note that, if maximum ratio combining (MRC) is considered as the precoding criterion for $\mathcal{W}_k$, then there is no difference between precoder feedback $\mathbf{w}_k^{\rm c}$ and CSI feedback $\hat{\mathbf{h}}_k$. By contrast, if zero-forcing (ZF) type criteria are used and the D2D CSI exchange has a much higher rate than the feedback to the BS, then the cooperative precoder feedback scheme $\mathcal{W}_k$ can exploit the advantage of both knowing the self channel $\mathbf{h}_k$ perfectly and knowing the channels from the other users more precisely. Such intuition will be analyzed in the next section.

III. Analysis of Cooperative Feedback for Identically Uncorrelated Channels

In this section, we focus on identically uncorrelated channels, where $l_k = 1$ and $\mathbf{R}_k = \mathbf{I}$; i.e., the entries of the channel vectors $\mathbf{h}_k$ are independent and identically distributed (i.i.d.) and follow $\mathcal{CN}(0,1)$. Since the users have identical CSI statistics, the same quantizer with the same rate $b_{kj} = B_{\rm c}$, $\forall k, j,\ k \neq j$, is used for the CSI exchange in the proposed scheme. We develop analytical results to compare the cooperative feedback scheme with the conventional CSI feedback scheme. As a compromise of mathematical tractability, we consider ZF-type precoding strategies.

A. The Schemes

The conventional CSI feedback scheme and the proposed cooperative precoder feedback scheme are specified as follows.

CSI feedback scheme: Random vector quantization (RVQ) is used for channel quantization and feedback, where each user k has a *channel codebook* $\mathcal{C}_k$ that contains $2^{B_{\rm f}}$ N_t-dimensional unit norm isotropic distributed vectors, and the channel $\mathbf{h}_k$ is quantized as $\hat{\mathbf{g}}_k = \arg\max_{\mathbf{u}\in\mathcal{C}_k} |\mathbf{h}_k^{\rm H}\mathbf{u}|$ and fed back to the BS. The BS computes the precoder $\mathbf{w}_k$ for user k as the normalized kth column of the precoding matrix
$$\mathbf{W} = \hat{\mathbf{G}}(\hat{\mathbf{G}}^{\rm H}\hat{\mathbf{G}})^{-1} \quad (2)$$
where $\hat{\mathbf{G}}$ is a $N_t \times K$ matrix with the kth column given by the quantized channel $\hat{\mathbf{g}}_k$. Note that, the channel magnitude $\|\mathbf{h}_k\|$ is not known by the BS.

Cooperative precoder feedback scheme: Each user has two codebooks: the *channel codebook* $\mathcal{C}_{kj}^{\rm c} = \mathcal{C}_k^{\rm c}$, $j \neq k$, that contains $2^{B_{\rm c}}$ N_t-dimensional unit norm isotropic distributed vectors for the CSI exchange, and the *precoder codebook* $\mathcal{C}_k^{\rm w}$ that contains $2^{B_{\rm f}}$ N_t-dimensional unit norm isotropic distributed vectors for the feedback to the BS. In the CSI exchange phase, the vector $\hat{\mathbf{h}}_k^{(j)} = \hat{\mathbf{h}}_k = \|\mathbf{h}_k\|\hat{\mathbf{g}}_k^{\rm c}$ is shared to all the users $j \neq k$, where $\hat{\mathbf{g}}_k^{\rm c} = \arg\max_{\mathbf{u}\in\mathcal{C}_k^{\rm c}} |\mathbf{h}_k^{\rm H}\mathbf{u}|$.[3] In the cooperative feedback phase, the vector that minimizes the interference leakage is fed back to the BS (as the counterpart of the ZF (2) in the CSI feedback scheme):
$$\mathbf{w}_k^{\rm c} = \arg\min_{\mathbf{w}\in\mathcal{C}_k^{\rm w}} \sum_{j\neq k} |\hat{\mathbf{h}}_j^{\rm H}\mathbf{w}|^2. \quad (3)$$

This section analyzes the performance in terms of interference leakage defined as
$$I_k = \rho \sum_{j\neq k} |\mathbf{h}_j^{\rm H}\mathbf{w}_k|^2$$

[3]The channel magnitude $\|\mathbf{h}_k\|$ is assumed to be shared among users with negligible distortion under additional $B_{\rm c}^{(0)}$ bits. Note that $B_{\rm c}^{(0)}$ needs not scale with N_t and will not affect the main insights of the results, and hence is ignored in this paper.

where $\rho = P/K$ denotes the power allocation. Before going through the derivations, we first state the main result of this section as follows.

Theorem 1 (Interference Leakage Upper Bound): Under large N_t and B_f, the average interference leakage under the precoder feedback scheme is roughly upper bounded by

$$\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_k\right\} \lesssim \rho 2^{-\frac{B_\mathrm{f}}{K-1}} + \rho(K-1)2^{-\frac{B_\mathrm{c}}{N_t-1}} \quad (4)$$

where the expectation $\mathbb{E}_{\mathcal{H},\mathcal{C}}\{\cdot\}$ is taken over the distributions of the channels and the codebooks.

It is known that the interference leakage of the CSI feedback scheme is lower bounded by $\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_k\right\} > \rho(K-1)2^{-\frac{B_\mathrm{f}}{N_t-1}}$ from [15]. Our result thus shows that for sufficiently large B_c for CSI exchange, the interference leakage from the precoder feedback scheme is dominated by the first term of (4), which is $K-1$ times lower than the CSI feedback scheme and decreases faster as B_f increases.

B. Characterization of the Interference Leakage

We first characterize the interference leakage in terms of the precoding vectors and the quantization errors. Without loss of generality, we focus on the performance of user 1.

Lemma 1 (Characterization of the Interference Leakage): The mean of the interference leakage $I_1 = \rho\sum_{j\neq 1}|\mathbf{h}_j^\mathrm{H}\mathbf{w}_1|^2$ can be characterized as

$$\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_1\right\} = \rho N_t \sum_{j\neq 1}\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{(1-Z_j)\left|\hat{\mathbf{g}}_j^\mathrm{H}\mathbf{w}_1\right|^2 + Z_j\left|\mathbf{s}_j^\mathrm{H}\mathbf{w}_1\right|^2\right\} \quad (5)$$

where $Z_j \triangleq 1-|\hat{\mathbf{g}}_j^\mathrm{H}\mathbf{g}_j|^2$ with $\mathbf{g}_j = \mathbf{h}_j/\|\mathbf{h}_j\|$ and $\mathbf{s}_j \triangleq (\mathbf{I}-\hat{\mathbf{g}}_j\hat{\mathbf{g}}_j^\mathrm{H})\mathbf{g}_j/\sqrt{Z_j}$.

Proof: Please refer to Appendix A for the proof. □

Note that the quantity Z_j captures the channel quantization error in magnitude, and $\mathbf{s}_j$ captures the difference in direction between the quantized vector $\hat{\mathbf{g}}_j$ and the true channel $\mathbf{g}_j$ in the (N_t-1)-dimensional space.[4] Thus, Lemma 1 illustrates that the average interference is a sum of the interference leakage due to precoding and the residual interference due to channel quantization errors.

Intuitive comparisons between precoder feedback and CSI feedback can be made from (5). In the CSI feedback scheme, the first term in (5) gives $\left|\hat{\mathbf{g}}_j^\mathrm{H}\mathbf{w}_1\right|^2 = 0$ due to the ZF precoding at BS. The second term characterizes the interference leakage due to quantization error, which is in terms of B_f. The results in [15] show that it is roughly $N_t\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{Z_j\left|\mathbf{s}_j^\mathrm{H}\mathbf{w}_1\right|^2\right\} \approx 2^{-\frac{B_\mathrm{f}}{N_t-1}}$.

In the precoder feedback scheme, the first term $\left|\hat{\mathbf{g}}_j^\mathrm{H}\mathbf{w}_1^\mathrm{c}\right|^2 \neq 0$, since $\mathbf{w}_1^\mathrm{c} \in \mathcal{C}_1^\mathrm{w}$ is chosen from a finite number of vectors. By contrast, the second term is affected by the quantization error in terms of B_c for CSI exchange. Usually B_c is large, and can be evaluated using existing results. Specifically, the channel quantization error bounds can be given as [15]

$$\frac{N_t-1}{N_t}2^{-\frac{B_\mathrm{c}}{N_t-1}} < \mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{Z_j\right\} < 2^{-\frac{B_\mathrm{c}}{N_t-1}} \quad (6)$$

[4]One can verify that $\mathbf{s}_j$ has unit norm and is orthogonal to $\hat{\mathbf{g}}_j$.

and $\mathbb{E}_{\mathcal{H},\mathcal{C}}\{\left|\mathbf{s}_j^\mathrm{H}\mathbf{w}_1^\mathrm{c}\right|^2\} = 1/(N_t-1)$. Moreover, $\left|\mathbf{s}_j^\mathrm{H}\mathbf{w}_1\right|^2$ is independent of Z_j.[5] As a result, $N_t\mathbb{E}_{\mathcal{H},\mathcal{C}}\{Z_j\left|\mathbf{s}_j^\mathrm{H}\mathbf{w}_1\right|^2\} < \frac{N_t}{N_t-1}2^{-B_\mathrm{c}/(N_t-1)}$.

In the following part, we focus on quantifying the first term in (5) for the interference leakage under the precoder feedback scheme with user cooperation.

C. Interference Upper Bound in a Two-user Case

In two-user case, the precoder $\mathbf{w}_1$ only depends on $\hat{\mathbf{g}}_2$, and the interference leakage from user 1 is just the interference at user 2. Using this insight, the interference upper bound under the precoder feedback scheme can be derived in the following theorem.

Theorem 2 (Interference Upper Bound for Two Users): The mean of the interference leakage $I_1^\mathrm{c} = \rho|\mathbf{h}_2^\mathrm{H}\mathbf{w}_1^\mathrm{c}|^2$ is upper bounded by

$$\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_1^\mathrm{c}\right\} \leq \frac{\rho N_t}{N_t-1}\left[2^{-B_\mathrm{f}} + \left(1-\frac{N_t-1}{N_t}2^{-B_\mathrm{f}}\right)2^{-\frac{B_\mathrm{c}}{N_t-1}}\right].$$

Proof: Please refer to Appendix B for the proof. □

The following corollary characterizes the case under perfect CSI exchange among users.

Corollary 1 (Interference Upper Bound under Perfect CSI Exchange): With perfect CSI exchange, i.e., $B_\mathrm{c} = \infty$, the interference is upper bounded by $\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_1^\mathrm{c}\right\} \leq \frac{\rho N_t}{N_t-1}2^{-B_\mathrm{f}}$.

As a comparison, the mean of the interference $I_1 = \rho|\mathbf{h}_2^\mathrm{H}\mathbf{w}_1|^2$ under the CSI feedback scheme with ZF precoding at the BS can be bounded as [15]:

$$\rho 2^{-\frac{B_\mathrm{f}}{N_t-1}} < \mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_1\right\} < \frac{\rho N_t}{N_t-1}2^{-\frac{B_\mathrm{f}}{N_t-1}}. \quad (7)$$

With imperfect CSI exchange, Theorem 2 shows that for $B_\mathrm{c} \gg B_\mathrm{f}$, the interference under the precoder feedback scheme is smaller, and it decreases faster than the CSI feedback scheme when increasing the number of feedback bits B_f. On the other hand, when B_c is small, the interference under precoder feedback scheme is dominated by the residual interference due to channel quantization errors for CSI exchange among users.

D. Interference Leakage in the K-user Case

In $K > 2$ user case, the precoding vector $\mathbf{w}_1^\mathrm{c}$ depends on more than one channel vectors, and hence the exact distribution of $\sum_{j\neq 1}|\hat{\mathbf{h}}_j^\mathrm{H}\mathbf{w}_1^\mathrm{c}|^2$ is difficult to obtain. We resolve this challenge by using large system approximations and the extreme value theory, assuming both N_t and 2^{B_f} are large.

Specifically, given a quantized channel realization $\{\hat{\mathbf{h}}_j\}_{j\neq 1}$ and a sequence of i.i.d. unit norm isotropic random vectors $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots$ independent of $\{\hat{\mathbf{h}}_j\}_{j\neq 1}$, we first approximate the random variables $\widetilde{Y}_i \triangleq \sum_{j\neq 1}|\hat{\mathbf{h}}_j^\mathrm{H}\tilde{\mathbf{w}}_i|^2$ as independent chi-square random variables (multiplied by a scale factor $\frac{1}{2}$) with

[5]To see these results, note that $\mathbf{s}_j$ follows isotropic distribution on the sphere in (N_t-1)-dimensional space, since both the vectors $\hat{\mathbf{g}}_k^\mathrm{c}$ in the channel codebook $\mathcal{C}_k^\mathrm{c}$ and the channel direction $\mathbf{g}_k$ are isotropically distributed in the N_t-dimensional space. As a result, $\left|\mathbf{s}_j^\mathrm{H}\mathbf{w}_k^\mathrm{c}\right|^2$ follows a beta distribution $\mathcal{B}(1, N_t-2)$ for any unit norm vector $\mathbf{w}_k^\mathrm{c}$ as studied in [15], and hence the mean is given by $1/(N_t-1)$.

degrees of freedom $2(K-1)$. Note that such approximation becomes exact in large N_t.

The following lemma gives the asymptotic distribution of $\widetilde{Y}_i$ under the large N_t regime.

Lemma 2 (Asymptotic Chi-square Distribution): Let $X_1, X_2, \ldots, X_N$ be a sequence of i.i.d. random variables that follows chi-square distribution $\chi^2(2(K-1))$. Then, $(\widetilde{Y}_1, \widetilde{Y}_2, \ldots, \widetilde{Y}_N)$ converges to $\frac{1}{2}(X_1, X_2, \ldots, X_N)$ in distribution, as $N_t \to \infty$.

Proof: Please refer to Appendix C for the proof. □

Lemma 2 shows that as N_t becomes large, the variables $\widetilde{Y}_i$ and $\widetilde{Y}_j$ tend to become independent and $\frac{1}{2}\chi^2(2(K-1))$ chi-square distributed.

Consider the minimum interference leakage precoding criterion in (3), and note that the precoding vector $\mathbf{w}_k^{\mathrm{c}}$ is chosen from a set of i.i.d. isotropic vectors $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots$ in $\mathcal{C}_k^{\mathrm{w}}$. Thus, the resultant interference leakage $\sum_{j\neq 1} |\hat{\mathbf{h}}_j^{\mathrm{H}} \mathbf{w}_k^{\mathrm{c}}|^2$ is approximately the minimum of $2^{B_{\mathrm{f}}}$ i.i.d. chi-square distributed (with a constant factor $\frac{1}{2}$) random variables $\widetilde{Y}_i = \sum_{j\neq 1} |\hat{\mathbf{h}}_j^{\mathrm{H}} \tilde{\mathbf{w}}_i|^2$, $i = 1, 2, \ldots, 2^{B_{\mathrm{f}}}$. As the codebook size $N = 2^{B_{\mathrm{f}}}$ is usually very large, one can apply extreme value theory to approximate the distribution of $\min_i \widetilde{Y}_i$ in order to yield simple expressions.

Let $\hat{I}_k \triangleq \sum_{j\neq k} |\hat{\mathbf{h}}_j^{\mathrm{H}} \mathbf{w}_k^{\mathrm{c}}|^2$, where $\mathbf{w}_k^{\mathrm{c}}$ is chosen from the precoder codebook $\mathcal{C}_k^{\mathrm{w}}$ under minimum interference leakage criterion (3). Thus, $\hat{I}_k = \min_i \widetilde{Y}_i$. Let $N = |\mathcal{C}_k^{\mathrm{w}}|$. The asymptotic property of $\hat{I}_k$ can be characterized in the following lemma.

Lemma 3 (Asymptotic Distribution of $\hat{I}_k$): The distribution of $\hat{I}_k$ satisfies

$$\lim_{N\to\infty} \lim_{N_t\to\infty} \mathbb{P}\left\{\hat{I}_k < \phi_N y\right\} = 1 - \exp(-y^{K-1}), \qquad x \geq 0$$

where

$$\phi_N = \sup\left\{x : \frac{1}{\Gamma(K-1)} \int_0^x t^{K-2} e^{-t} dt \leq \frac{1}{N}\right\} \tag{8}$$

in which $\Gamma(x)$ denotes the Gamma function. Moreover, for small K, ϕ_N can be approximated by

$$\phi_N \approx \Gamma(K)^{-\frac{1}{K-1}} N^{-\frac{1}{K-1}}. \tag{9}$$

Proof: Please refer to Appendix D for the proof. □

Lemma 3 suggests that for large $N = 2^{B_{\mathrm{f}}}$ and large N_t, the interference leakage $\hat{I}_k$ due to finite precoding can be approximated by a random variable $\phi_N W_{K-1}$ in distribution, where W_{K-1} is Weibull distributed with cumulative distribution function (CDF) given by $f_W(x; K-1) = 1 - \exp(-x^{K-1})$, $x \geq 0$, and mean $\mathbb{E}\{W_{K-1}\} = \Gamma\left(\frac{K}{K-1}\right)$.

With these results, the mean interference leakage under the precoder feedback scheme can be derived in the following theorem.

Theorem 3 (Interference Leakage for K Users): The mean of the interference leakage $I_k^{\mathrm{c}} = \rho \sum_{j\neq k} |\mathbf{h}_j^{\mathrm{H}} \mathbf{w}_k^{\mathrm{c}}|^2$ under K-user networks can be approximated by

$$\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_k^{\mathrm{c}}\right\} \approx \rho \Gamma\left(\frac{K}{K-1}\right) \phi_N + \rho \left[\frac{N_t(K-1)}{N_t - 1} - \frac{N_t - 1}{N_t} \Gamma\left(\frac{K}{K-1}\right) \phi_N\right] 2^{-\frac{B_{\mathrm{c}}}{N_t - 1}} \tag{10}$$

where ϕ_N is given in (8) with $N = 2^{B_{\mathrm{f}}}$. In addition, for small K,

$$\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_k^{\mathrm{c}}\right\} \approx \rho \Phi(K) 2^{-\frac{B_{\mathrm{f}}}{K-1}} + \rho \left[\frac{N_t(K-1)}{N_t - 1} - \frac{N_t - 1}{N_t} \Phi(K) 2^{-\frac{B_{\mathrm{f}}}{K-1}}\right] 2^{-\frac{B_{\mathrm{c}}}{N_t - 1}} \tag{11}$$

where $\Phi(K) \triangleq \Gamma(\frac{K}{K-1}) \Gamma(K)^{-\frac{1}{K-1}}$.

Proof: Using the results in Lemma 1 and 3, the derivation is similar to Theorem 2, and is omitted here due to limited space. □

One can numerically verify that the term $\Phi(K)$ is decreasing in K and $\Phi(K) \leq 1$ for $K \geq 2$. Therefore, under sufficiently large B_{c} and N_t, the interference leakage $\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_k^{\mathrm{c}}\right\}$ is roughly upper bounded by $\rho 2^{-\frac{B_{\mathrm{f}}}{K-1}} + \rho(K-1) 2^{-\frac{B_{\mathrm{c}}}{N_t - 1}}$, which is significantly smaller than that of the CSI feedback scheme $\rho(K-1) 2^{-\frac{B_{\mathrm{f}}}{N_t - 1}}$. On the other hand, in the undesired small B_{c} regimes, the second term in (11) dominates, which represents the residual interference due to poor quantization for CSI exchange among users.

IV. Adaptive CSI Exchange for Non-identical Channels

In this section, we study the case of non-identical channels, where users may have different path loss l_k and different channel covariance matrices $\mathbf{R}_k$. In this scenario, it is not efficient to distribute equal bits to the users for CSI exchange. The intuitions are as follows. First, some users may be in the interference limited region and require the other users to know their channels for interference aware precoding, whereas some other users may be in the noise limited region and inter-user interference is not an essential issue for them. Second, when two users have non-overlapping signal subspaces, they do not need to exchange the CSI, because there is no interference for each other even under MRC precoding. Therefore, the users should have different CSI exchange strategies according to the global CSI statistics $\{l_k, \mathbf{R}_k\}$ and the whole D2D resources B_{tot} bits should be smartly partitioned over all the user pairs.

We first specify the precoding strategy for cooperative precoder feedback scheme. Then, we elaborate the proposed CSI exchange strategy and analyze the interference leakage for the cooperative precoder feedback scheme. Based on this, we derive the optimal bit partition for CSI exchange.

A. Precoding Strategy

Consider the following precoder codebook

$$\mathcal{C}_k^{\mathrm{w}} = \left\{\mathbf{u}_i : \mathbf{u}_i = \mathbf{R}_k^{\frac{1}{2}} \boldsymbol{\xi}_i / \|\mathbf{R}_k^{\frac{1}{2}} \boldsymbol{\xi}_i\|_2, \, i = 1, 2, \ldots, 2^{B_{\mathrm{f}}}\right\} \tag{12}$$

where $\boldsymbol{\xi}_i$ are random vectors following complex Gaussian distribution $\mathcal{CN}(\mathbf{0}, \mathbf{I})$.

With the imperfect global CSI $\hat{\mathbf{H}}_k$, the signal-to-interference-leakage-and-noise-ratio (SLNR) of user k using precoder $\mathbf{w} \in \mathcal{C}_k^{\mathrm{w}}$ can be computed as follows

$$\mathbb{E}\left\{\frac{\left|\mathbf{h}_k^{\mathrm{H}} \mathbf{w}\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_j^{\mathrm{H}} \mathbf{w}\right|^2 + \alpha} \middle| \hat{\mathbf{H}}_k\right\} \geq \frac{\left|\mathbf{h}_k^{\mathrm{H}} \mathbf{w}\right|^2}{\sum_{j\neq k} \mathbb{E}\left\{|\mathbf{h}_j^{\mathrm{H}} \mathbf{w}|^2 \,|\hat{\mathbf{H}}_k\right\} + \alpha} \tag{13}$$

where $\alpha = K/P$ and Jesens's inequality $\mathbb{E}\{f(x)\} \geq f(\mathbb{E}\{x\})$ is used for the convex function $f(x) = 1/(x+\alpha)$. From the CSI model (1), we have $\mathbf{h}_j = \hat{\mathbf{h}}_j^{(k)} + (\mathbf{h}_j - \hat{\mathbf{h}}_j^{(k)})$, and

$$\mathbb{E}\{|\mathbf{h}_j^{\mathrm{H}}\mathbf{w}|^2 \,|\hat{\mathbf{H}}_k\} = |\hat{\mathbf{h}}_j^{(k)\mathrm{H}}\mathbf{w}|^2 + \mathbf{w}^{\mathrm{H}}\mathbf{Q}_{jk}\mathbf{w}$$

where

$$\mathbf{Q}_{jk} \triangleq \mathbb{E}\{(\mathbf{h}_j - \hat{\mathbf{h}}_j^{(k)})(\mathbf{h}_j - \hat{\mathbf{h}}_j^{(k)})^{\mathrm{H}}\} \tag{14}$$

is the error covariance of the CSI exchanged via D2D. Note that the expectation in (13) is to average over errors due to imperfect CSI exchange.

As a result, we propose to choose a precoding vector $\mathbf{w}_k^{\mathrm{c}} = \mathcal{W}_k(\hat{\mathbf{H}}_k)$ to maximize the expected SLNR lower bound (13) as follows

$$\underset{\mathbf{w}\in\mathcal{C}_k^{\mathrm{w}}}{\mathrm{maximize}} \quad \frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}|^2}{\sum_{j\neq k}|\hat{\mathbf{h}}_j^{(k)\mathrm{H}}\mathbf{w}|^2 + \mathbf{w}^{\mathrm{H}}\sum_{j\neq k}\mathbf{Q}_{jk}\mathbf{w} + \alpha}. \tag{15}$$

The motivation to use SLNR precoder is that first, it can be computed in a distributive way, and second, the SLNR precoding has been shown to achieve good performance in multiuser MIMO systems from low to high signal-to-noise ratio (SNR) [33]–[35]. In particular, there is a strong relation between SLNR precoding and minimum mean square error (MMSE) precoding.

Remark 1 (Connection between SLNR Precoding and MMSE Precoding): Consider precoding strategies in the continuous domain (i.e., without constrained in the precoder codebook $\mathcal{C}_k^{\mathrm{w}}$). The transmit precoding vector that satisfies MMSE criteria is given by

$$\mathbf{w}_k^{\mathrm{MMSE}} = \sqrt{\Psi_k}\left(\hat{\mathbf{H}}_k\hat{\mathbf{H}}_k^{\mathrm{H}} + \alpha\mathbf{I}\right)^{-1}\mathbf{h}_k \tag{16}$$

where Ψ_k is a normalizing factor such that $\|\widetilde{\mathbf{w}}_{k,\mathrm{MMSE}}\|^2 = 1$. On the other hand, the SLNR precoding vector is given by

$$\mathbf{w}_k^{\mathrm{SLNR}} = \arg\max_{\|\mathbf{w}\|^2=1} \frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}|^2}{\sum_{j\neq k}|\hat{\mathbf{h}}_j^{(k)\mathrm{H}}\mathbf{w}|^2 + \alpha}. \tag{17}$$

It was shown in [35] that the MMSE precoder in (16) is equivalent to the SLNR precoder in (17) up to a complex scaling, i.e., $\mathbf{w}_k^{\mathrm{MMSE}} = c_k\mathbf{w}_k^{\mathrm{SLNR}}$.

B. CSI Exchange Strategy

The proposed CSI exchange strategy consists of two components, namely, subspace projection for dimension reduction, and D2D quantizer for bit partition among different user pairs.

1) Subspace Projection: We propose a channel quantization method for CSI exchange based on *signal subspace projection*. The strategy consists of two steps.

- Subspace projection: To share the channel $\mathbf{h}_k$ to user j, user k first computes the partial channel

$$\mathbf{g}_k^{(j)} = \frac{1}{\sqrt{l_k}}\mathbf{U}_j^{\mathrm{H}}\mathbf{h}_k \tag{18}$$

  where $\mathbf{U}_j$ is a $N_t \times \bar{M}_j$ matrix that contains the $\bar{M}_j$ dominant eigenvectors of the covariance matrix $\mathbf{R}_j$ of user j.
- Quantization: The partial channel $\mathbf{g}_k^{(j)}$ is quantized into $\hat{\mathbf{g}}_k^{(j)}$ using b_{kj} bits and transmitted to user j.

User j obtains the channel from user k as $\hat{\mathbf{h}}_k^{(j)} = \sqrt{l_k}\mathbf{U}_j\hat{\mathbf{g}}_k^{(j)}$.

Remark 2 (Intuitive Interpretation): The intuition of the subspace projection is that, only the portion of the channel that lies in the overlapping signal subspace needs to be exchanged. To see this, rewrite the channel of user k as

$$\begin{aligned}\mathbf{h}_k &= \mathbf{U}_j\mathbf{U}_j^{\mathrm{H}}\mathbf{h}_k + (\mathbf{I} - \mathbf{U}_j\mathbf{U}_j^{\mathrm{H}})\mathbf{h}_k \\ &= \mathbf{h}_k^{(j)} + \mathbf{h}_k^{(j)\perp}\end{aligned} \tag{19}$$

where $\mathbf{h}_k^{(j)}$, which can be written as $\mathbf{h}_k^{(j)} = \sqrt{l_k}\mathbf{U}_j\mathbf{g}_k^{(j)}$, is the portion of $\mathbf{h}_k$ that lies in the overlapping signal subspace of users k and j, whereas, $\mathbf{h}_k^{(j)\perp}$ is orthogonal to the overlapping signal subspace. From the construction of precoder codebook $\mathcal{C}_j^{\mathrm{w}}$ in (12), the precoder $\mathbf{w}_j^{\mathrm{c}}$ lies in the subspace spanned by $\mathbf{U}_j$. As a result, $|(\mathbf{h}_k^{(j)\perp})^{\mathrm{H}}\mathbf{w}_j^{\mathrm{c}}| = 0$ almost surely, and hence there is no need to transmit $\mathbf{h}_k^{(j)\perp}$ to user j.

2) D2D Quantizer: Note that in the conventional CSI feedback scheme, the CSI is used for both signal enhancement and interference mitigation, whereas in the proposed precoder feedback scheme (15), the CSI exchanged among users is for interference mitigation only. As a result, not all the users require the same level of CSI quality, depending on the propagation scenarios such as signal subspace and path loss.

The concept of D2D quantizer for CSI exchange among users is highlighted as follows.

Definition 1 (D2D Quantizer): A D2D quantizer $\mathcal{Q}(\{b_{kj}\})$ with total bits B_{tot} consists of bit partition $\{b_{kj} : \sum_{k=1}^{K}\sum_{j\neq k} b_{kj} = B_{\mathrm{tot}}\}$ and a set of individual quantizers $\mathcal{Q}_{kj}$ with rate b_{kj} that map the partial channel $\mathbf{g}_k^{(j)}$ to $\hat{\mathbf{g}}_k^{(j)}$.

There are many techniques to design the quantizers $\mathcal{Q}_{kj}$. For example, for small number of bits b_{kj}, codebook based vector quantization techniques can be used [10]–[14]. Here, we choose entropy-coded scalar quantization for elaboration [36], because it is easier to scale to moderate or large number of bits b_{kj} for the scenario of CSI exchange via D2D.

3) CSI Exchange using Entropy-coded Scalar Quantization: Consider an entropy-coded scalar quantizer [36] designed as follows.

First, Karhunen-Loève Transform (KLT) is applied to de-correlate the entries of the vector $\mathbf{g}_k^{(j)}$. Let $\mathbf{G}_{kj} \triangleq \mathbb{E}\{\mathbf{g}_k^{(j)}\mathbf{g}_k^{(j)\mathrm{H}}\}$ be the covariance matrix of the partial channel $\mathbf{g}_k^{(j)}$ and denote the eigen decomposition of $\mathbf{G}_{kj}$ as $\mathbf{G}_{kj} = \mathbf{U}_{kj}^{\mathrm{H}}\mathbf{\Lambda}_{kj}\mathbf{U}_{kj}$, where $\mathbf{\Lambda}_{kj} = \mathrm{diag}(\lambda_{kj}^{(1)}, \lambda_{kj}^{(2)}, \ldots, \lambda_{kj}^{(M_{kj})})$ is a diagonal matrix that contains the eigenvalues $\{\lambda_{kj}^{(i)}\}$ of $\mathbf{G}_{kj}$ in descending order. The KLT of $\mathbf{g}_k^{(j)}$ is given by

$$\mathbf{q}_k^{(j)} = \mathbf{U}_{kj}^{\mathrm{H}}\mathbf{g}_k^{(j)}. \tag{20}$$

Note that there is a dimension reduction $M_{kj} < \bar{M}_j$ when the subspaces of user k and j are only partially overlapped.

With the KLT and since the channel vector $\mathbf{h}_k$ is $\mathcal{CN}(\mathbf{0}, l_k\mathbf{R}_k)$ distributed, the ith element of $\mathbf{q}_k^{(j)}$ follows $\mathcal{CN}(0, \lambda_{kj}^{(i)})$, and is uncorrelated with the other elements of $\mathbf{q}_k^{(j)}$. Then, a scalar quantizer is designed to quantize each element of $\mathbf{q}_k^{(j)}$, and at the same time, lossless code (such as Hamming code) is applied to encode the output of the quantizer, such that the average output bit rate approaches to the entropy of the quantizer, where the entropy is constrained to be b_{kj}.

Finally, at user j, the channel of user k is reconstructed as $\hat{\mathbf{h}}_k^{(j)} = \sqrt{l_k}\mathbf{U}_j\hat{\mathbf{g}}_k^{(j)} = \sqrt{l_k}\mathbf{U}_j\mathbf{U}_{kj}\hat{\mathbf{q}}_k^{(j)}$.

Define the distortion of $\hat{\mathbf{g}}_k^{(j)}$ as the squared error given by

$$D_{kj} \triangleq \mathbb{E}\left\{\|\mathbf{g}_k^{(j)} - \hat{\mathbf{g}}_k^{(j)}\|_2^2\right\}. \quad (21)$$

The distortion-rate function $D_{kj}(b_{kj})$ is defined as the theoretical minimum distortion D_{kj} under b_{kj} bits. As a direct application of Shannon's distortion-rate theory [37, Theorem 13.3.3], the distortion-rate function for the above quantizer can be given in the following lemma.

Lemma 4 (Distortion-rate): The distortion-rate function $D_{kj}(b_{kj})$ is given by

$$D_{kj}(b_{kj}) = m_{kj}^* \left(\prod_{i=1}^{m_{kj}^*} \lambda_{kj}^{(i)}\right)^{\frac{1}{m_{kj}^*}} 2^{-\frac{b_{kj}}{m_{kj}^*}} + \sum_{i=m_{kj}^*+1}^{M_{kj}} \lambda_{kj}^{(i)} \quad (22)$$

where $m_{kj}^* \leq M_{kj}$ is a positive integer such that $\sum_{i=1}^{m_{kj}^*} r_i(m_{kj}^*) = b_{kj}$, in which $r_i(m) = \max\left\{0, \frac{b_{kj}}{m} + \log_2\left[\lambda_{kj}^{(i)}/(\prod_{i=1}^{m}\lambda_{kj}^{(i)})^{\frac{1}{m}}\right]\right\}$.

Remark 3 (Achievability): The distortion-rate function $D_{kj}(b_{kj})$ in (22) can be roughly achieved by applying infinite level uniform scalar quantizer [38] with reverse water-filling bit allocation to distribute b_{kj} bits over the elements of $\mathbf{q}_k^{(j)}$ [37, Theorem 13.3.3], and at the same time, encoding the output of the quantizer using lossless codes (such as Hamming code). Note that the operational distortion-rate $\hat{D}_{kj}(b_{kj})$ under such method asymptotically approaches to the Shannon's distortion-rate function (22) in low resolution regime (small b_{kj}) [38]. In high resolution regime, it requires additional 0.25 bit per real dimension to achieve the same distortion as $D_{kj}(b_{kj})$. Nevertheless, it is still insightful to apply $D_{kj}(b_{kj})$ in (22) to analyze the performance and optimize b_{kj} in the remaining part of the paper.

C. Bit Partition for CSI Exchange

Intuitively, the bit partition for CSI exchange should mainly depend on the channel statistics $\{l_k, \mathbf{R}_k\}$ and the quantizers $\mathcal{Q}_{kj}$, but should not be quite affected by the number of feedback bits B_{f}. To make the analysis tractable and isolate the impact of B_{f}, the concept of *virtual SLNR* is introduced as follows.

Definition 2 (Operational SLNR): The operational SLNR of user k is defined as

$$\gamma_k(\mathbf{H}, \mathcal{Q}, \mathcal{C}_k^{\text{w}}) \triangleq \frac{|\mathbf{h}_k^{\text{H}}\mathbf{w}_k^{\text{c}}|^2}{\sum_{j\neq k}|\mathbf{h}_j^{\text{H}}\mathbf{w}_k^{\text{c}}|^2 + \alpha}$$

where $\alpha > 0$ is some regularization parameter, and $\mathbf{w}_k^{\text{c}} = \mathcal{W}_k(\hat{\mathbf{H}}_k)$ is the precoder from (15) that is based on the imperfect global CSI $\hat{\mathbf{H}}_k$ (depending on the CSI exchange quantizer $\mathcal{Q}$) and the precoder codebook $\mathcal{C}_k^{\text{w}}(B_{\text{f}})$, which contains $2^{B_{\text{f}}}$ precoding vectors.

Definition 3 (Virtual SLNR $\bar{\Gamma}_k$): Given the bit partition $\{b_{kj}\}$, SLNR Γ_k is achievable if there exists a D2D quantizer $\mathcal{Q}(\{b_{kj}\})$ and a sequence of precoder codebooks $\mathcal{C}_k^{\text{w}}(B_{\text{f}})$ such that $\lim_{B_{\text{f}}\to\infty}\mathbb{E}\{\gamma_k(\mathbf{H}, \mathcal{Q}, \mathcal{C}_k^{\text{w}})\} \geq \Gamma_k$. The virtual SLNR $\bar{\Gamma}_k(\{b_{kj}\})$ is the supremum of the achievable SLNR Γ_k.

The virtual SLNR $\bar{\Gamma}_k(\{b_{kj}\})$ is a function to characterize the theoretical performance of the bit partition $\{b_{kj}\}$ for CSI exchange. It isolates the impacts from the precoder codebook $\mathcal{C}_k^{\text{w}}$ and the parameter B_{f}. Ideally, the virtual SLNR $\bar{\Gamma}_k(\{b_{kj}\})$ can be achieved by SLNR precoding in the continuous domain $\|\mathbf{w}_k\| = 1$ (as in (17)) and optimal quantizers $\mathcal{Q}_{kj}$ for CSI exchange. As a result, the virtual SLNR $\bar{\Gamma}_k(\{b_{kj}\})$ serves as a good performance metric for bit partition.

Specifically, the bit partition that maximizes the virtual SLNR is formulated as follows

$$\begin{aligned} \underset{\{b_{kj}\geq 0\}}{\text{maximize}} \quad & \sum_{k=1}^{K}\log(\bar{\Gamma}_k(\{b_{kj}\})) \quad (23) \\ \text{subject to} \quad & \sum_{k=1}^{K}\sum_{j\neq k} b_{kj} = B_{\text{tot}} \end{aligned}$$

where the log function in the objective is to impose proportional fairness among users.

1) Virtual SLNR Lower Bound: The explicit expression of virtual SLNR in (23) is difficult to obtain. Instead, a lower bound can be derived as follows.

We first study the model of partial CSI $\hat{\mathbf{g}}_k^{(j)}$.

Lemma 5 (Distortion-rate under High Resolution CSI Exchange): For sufficiently large b_{kj}, the distortion-rate covariance $\mathbf{Q}_{kj}$ in (14) satisfies

$$\begin{aligned} \mathbf{Q}_{kj} = l_k M_{kj}\left(\prod_{i=1}^{M_{kj}}\lambda_{kj}^{(i)}\right)^{\frac{1}{M_{kj}}} 2^{-\frac{b_{kj}}{M_{kj}}}\mathbf{U}_j\mathbf{U}_{kj}\mathbf{U}_{kj}^{\text{H}}\mathbf{U}_j^{\text{H}} \\ + (\mathbf{I} - \mathbf{U}_j\mathbf{U}_j^{\text{H}})\mathbf{R}_k(\mathbf{I} - \mathbf{U}_j\mathbf{U}_j^{\text{H}})^{\text{H}} \end{aligned}$$

where the unitary matrices $\mathbf{U}_{kj}$ and $\mathbf{U}_j$ are defined in (20) and (18), respectively.

Proof: Please refer to Appendix E. □

The result in Lemma 5 is based on the Shannon distortion-rate function (22), which can be approached by entropy-coded scalar quantizer under high resolution, i.e., large b_{jk}. With such result, the lower bound of the virtual SLNR can be derived as follows.

Lemma 6 (Virtual SLNR Lower Bound): For sufficiently large b_{kj}, the virtual SLNR $\bar{\Gamma}_k$ is lower bounded by

$$\bar{\Gamma}_k(\{b_{kj}\}) \geq l_k \sum_{i=K}^{N_t}\lambda_k^{(i)}\left[\sum_{j\neq k} l_j\left(\prod_{m=1}^{M_{jk}}\lambda_{jk}^{(m)}\right)^{\frac{1}{M_{jk}}} 2^{-\frac{b_{jk}}{M_{jk}}} + \alpha\right]^{-1}. \quad (24)$$

Proof: Please refer to Appendix F. □

Remark 4 (Tightness of the lower bound): The lower bound (24) uses the Shannon distortion-rate function to approximate the virtual SLNR. Hence, the implementation of the actual quantizer affects the tightness of the bound.

2) Optimal Bit Partition: With the explicit expression on the virtual SLNR lower bound, the bit partition problem via SLNR maximization (23) can be reformulated as maximizing the virtual SLNR lower bound (24).

Note that since

$$\log\left(l_k\sum_{i=K}^{N_t}\lambda_k^{(i)}\left(\sum_{j\neq k}\omega_{jk}2^{-\frac{b_{jk}}{M_{jk}}}+\alpha\right)^{-1}\right) \tag{25}$$

$$=\log\left(l_k\sum_{i=K}^{N_t}\lambda_k^{(i)}\right)-\log\left(\sum_{j\neq k}\omega_{jk}2^{-\frac{b_{jk}}{M_{jk}}}+\alpha\right) \tag{26}$$

where $\omega_{jk} \triangleq l_j\left(\prod_{i=1}^{M_{jk}}\lambda_{jk}^{(i)}\right)^{\frac{1}{M_{jk}}}$, maximizing (25) over $\{b_{kj}\}$ is equivalent to minimizing the second term of (26). Specifically, the bit partition problem can be reformulated as follows

$$\begin{aligned}\underset{\{b_{kj}\geq 0\}}{\text{minimize}} \quad & \sum_{k=1}^{K}\log\left(\sum_{j\neq k}\omega_{jk}2^{-\frac{b_{jk}}{M_{jk}}}+\alpha\right) \\ \text{subject to} \quad & \sum_{k=1}^{K}\sum_{j\neq k}b_{kj}=B_{\text{tot}}.\end{aligned} \tag{27}$$

The minimization problem (27) is convex and the optimal solution can be obtained.

Let $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{k-1,k}, x_{k+1,k}, \ldots, x_{Kk})^{\mathrm{T}}$ and $\boldsymbol{\omega}_k = (\omega_{1k}, \omega_{2k}, \ldots, \omega_{k-1,k}, \omega_{k+1,k}, \ldots, \omega_{Kk})^{\mathrm{T}}$ be two vectors each with $K-1$ entries.

Theorem 4 (Optimal Bit Partition): The optimal bit partition that minimizes (27) is given by

$$b_{jk} = [-M_{jk}\log_2 x_{jk}]^+ \tag{28}$$

where $[x]^+ = \max\{0, x\}$, and x_{jk} is an entry in vector $\mathbf{x}_k$ given by

$$\mathbf{x}_k = \mu\alpha(\mathbf{A}_k - \mu\mathbf{1}\boldsymbol{\omega}_k^{\mathrm{T}})^{-1}\mathbf{1} \tag{29}$$

in which,

$$\mathbf{A}_k = \ln 2\cdot\mathrm{diag}(M_{1k}^{-1}, M_{2k}^{-1}, \ldots, M_{k-1,k}^{-1}, M_{k+1,k}^{-1}, \ldots, M_{Kk}^{-1})$$

the parameter μ is a non-negative variable chosen such that $\sum_{k=1}^{K}\sum_{j\neq k} b_{kj}(\mu) = B_{\text{tot}}$, and $\mathbf{1}$ is a $K-1$ dimensional column vector with all the entries being 1's,

Proof: Please refer to Appendix G. □

Note that, since the problem is convex, the parameter μ can be found using bisection search, which converges very fast.

The results in Theorem 4 suggests that the optimal bit partition for CSI exchange varies according to the path loss l_k, the dimension M_{jk} of the interference subspace between user k and j, as well as the eigenvalues of the covariance $\widetilde{\mathbf{R}}_{jk}$ of the overlapping subspace. First, from the objective (27), more bits b_{jk} should be allocated to a high SNR user j (i.e., large l_j), since user j is in the interference-limited regime and sensitive to the interference-aware precoding based on the CSI informed to the nearby users. Second, from the solution (28), more bits should be allocated to users which share a large dimensional subspace (i.e. large M_{jk}).

Remark 5 (Summary on signaling required for the adaptive CSI exchange): The BS collects the global CSI statistics $\{l_k, \mathbf{R}_k\}$ and computes the solution $\{b_{kj}\}$. Then the bit allocation b_{kj} and the subspace projectors $\mathbf{U}_j\mathbf{U}_{kj}$ are informed to all the users for adaptive CSI exchange. Note that these steps only depend on long-term CSI statistics, and hence the overhead is considered to be negligible in the long run.

V. Numerical Results

In this section, we evaluate the performance of the precoding feedback scheme with adaptive CSI exchange when users have different CSI statistics.[6]

Consider a single cell downlink massive MIMO system with $N_t = 60$ antennas at the BS serving $K = 2$ single antenna users. The noise variance is normalized to 1. The one-ring model [39] on uniform linear antenna array (ULA) is used for the channel modeling. The angular spread is 15 degrees and the power angular spectrum density follows a truncated Gaussian distribution centered at the mean azimuth direction of the user. A two-layer precoding structure [22], [23] is used and the precoding vector is given by $\mathbf{v}_k = \mathbf{\Phi}\mathbf{w}_k$, where $\mathbf{\Phi}$ is a $N_t \times M$ pre-beamforming matrix that contains $M = 20$ dominant eigenvectors of the joint CSI subspace characterized by $\sum_k \mathbb{E}\{\tilde{\mathbf{h}}_k\tilde{\mathbf{h}}_k^{\mathrm{H}}\}$, in which $\tilde{\mathbf{h}}_k^{\mathrm{H}}$ denotes the 1×60 channel vector of user k. The results in this paper can be directly applied by simply focusing on the equivalent channel $\mathbf{h}_k \triangleq \mathbf{\Phi}^{\mathrm{H}}\tilde{\mathbf{h}}_k$, which is still a large vector with 20 complex elements.

To incorporate the penalties due to the CSI sharing delays, consider the autoregressive time-variation model [40] for CSI $\mathbf{h}_k = \theta\mathbf{h}_k^{\mathrm{d}} + \sqrt{1-\theta^2}\boldsymbol{\xi}_k$, where $\theta = J_0(2\pi f_{\mathrm{d}}\tau_{\mathrm{d}})$ is the correlation coefficient, $J_0(\cdot)$ is the zeroth order Bessel function, $f_{\mathrm{d}} = \frac{v}{c}f_{\mathrm{c}}$ is the maximum doppler frequency under carrier frequency $f_{\mathrm{c}} = 2$ GHz, propagation speed $c = 3\times 10^8$ m/s, and user mobility speed $v = 5$ km/h, $\tau_{\mathrm{d}} = 10$ ms is the delay for CSI exchange, and $\boldsymbol{\xi}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. In the proposed schemes, each user k quantizes the delay CSI $\mathbf{h}_k^{\mathrm{d}}$ to $\hat{\mathbf{h}}_k^{(j)}$ and transmit to user j via D2D.

Each user has $B_{\mathrm{f}} = 6$ bits to feedback the precoder or the CSI to the BS, and the two users have in total $B_{\text{tot}} = 80$ bits for CSI exchange in the precoder feedback schemes.

The following CSI exchange, feedback, and precoding schemes are evaluated

- **Baseline 1 (CSI Feedback):** MMSE precoding is computed by the BS according to the CSI feedback from each user in B_{f} bits.
- **Baseline 2 (Precoder feedback with naive CSI exchange):** The CSI is quantized and exchanged according to each user's own CSI statistics, using $B_{\text{tot}}/2$ bits for each user. The precoder is computed according to (15) and fed back to the BS.
- **Proposed (Precoder feedback with adaptive CSI exchange):** The CSI is quantized and exchanged according to the proposed strategy in Section IV-B with adaptive bit partition for each user as in Section IV-C. The precoder is computed according to (15) and fed back to the BS.

A. Heterogeneous Path Loss

Consider the two users are near to each other and therefore they share the same signal subspace. However user 2 suffers from larger path loss due to additional blockage.[7] As a result, the two user have the same signal subspace, but user 2 suffers from larger path loss.

[6]The numerical results for the analysis in identically uncorrelated channels can be found in [32].

[7]For example, user 1 is outdoor and user 2 is indoor.

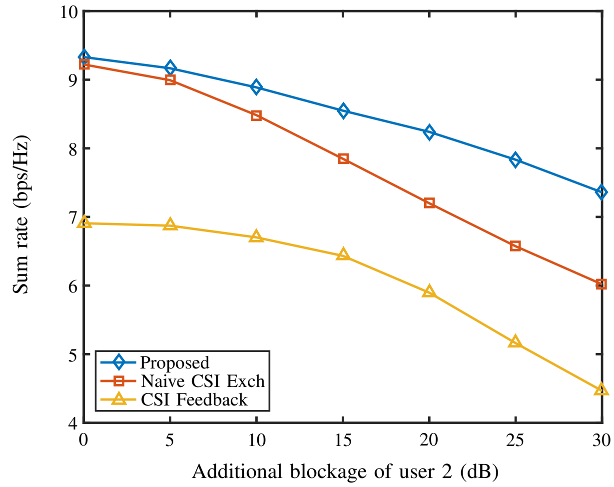

Figure 2. Sum rate versus additional blockage of user 2 under total transmission power $P = 20$ dB.

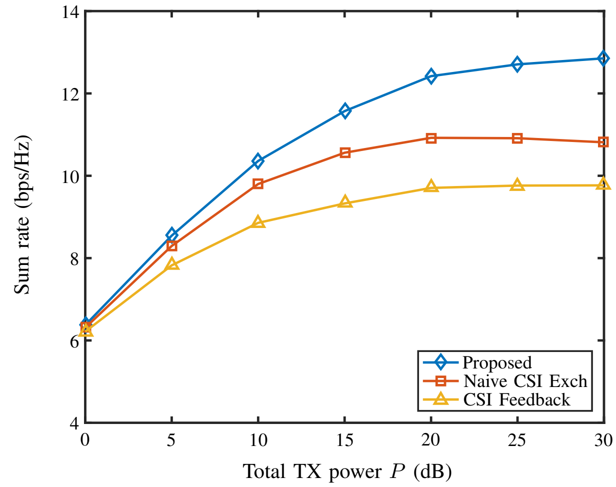

Figure 3. Sum rate versus the total transmission power.

Table II
BIT PARTITION FOR ADAPTIVE CSI EXCHANGE ACCORDING TO ADDITIONAL BLOCKAGE OF USER 2.

Blockage of user 2 (dB)	0	5	10	15	20	25
Bits for user 1 to quantize $\mathbf{h}_1^{(2)}$	40	49	58	67	76	80
Bits for user 2 to quantize $\mathbf{h}_2^{(1)}$	40	31	22	13	4	0

Fig. 2 shows the sum rate versus additional blockage of user 2 under total transmission power $P = 20$ dB. Specifically, the path loss of user 1 is normalized to 1, and the path loss of user 2 is equal to the blockage. First, both precoder feedback schemes significantly outperform the CSI feedback scheme. Second, the proposed precoder feedback with adaptive CSI exchange outperforms the naive CSI exchange scheme. This is because, user 2 is in the noise limited region, and hence it is not necessary for user 2 to inform its CSI $\mathbf{h}_2$ to user 1 for interference mitigation. On the other hand, user 1 wishes user 2 to know its CSI $\mathbf{h}_1$ for interference aware precoding, since user 1 is in interference limited region. Therefore, equal bit partition in the naive CSI exchange scheme is not efficient. The bit partition results for the proposed adaptive CSI exchange scheme is summarized in Table II.

B. Heterogeneous Signal Subspace

Consider that the two users have the same path loss (normalized to 1), but the users are separated by 10 meters and away from the BS by 60 meters. As a result, they have different signal subspace due to the limited angular spread.

1) Sum Rate Performance: Fig. 3 shows the sum rate versus the total transmission power. First, both precoder feedback schemes outperform the CSI feedback scheme. Second, the proposed precoder feedback with adaptive CSI exchange outperforms the naive CSI exchange scheme, because the proposed scheme quantizes the CSI using the statistics of both users. Specifically, it only quantizes the portion of CSI that lies in the overlapping signal subspace of the two users, and hence the quantization is more efficient.

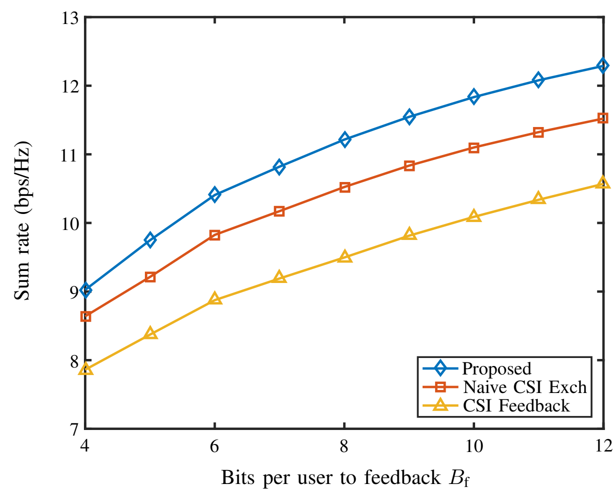

Figure 4. Sum rate versus the number of bits per user B_f for the feedback to the BS.

2) Feedback Saving: Fig. 4 demonstrates the sum rate versus the number of bits B_f per user for the feedback to the BS under total transmission power $P = 10$ dB. Both precoder feedback schemes outperform the CSI feedback scheme under $B_\mathrm{f} = 4$ to 12 feedback bits. In particular, the proposed scheme saves almost half of the bits for the feedback to the BS under similar sum rate performance as the CSI feedback scheme.

3) D2D Signaling Saving: Fig. 5 shows the sum rate versus total number of bits B_tot for CSI exchange under total transmission power $P = 10$ dB. The CSI feedback scheme is not affected by B_tot. The result demonstrates that when there are sufficient number of bits for CSI exchange, precoder feedback is preferred over CSI feedback. Under limited feedback to the BS and limited D2D signaling, the proposed scheme saves one third to almost half of the bits for CSI exchange as compared to the naive CSI exchange scheme.

4) Robustness under Delays for CSI Exchange: Fig. 6 shows the sum rate performance versus various delays for CSI exchange over D2D, where the total transmission power

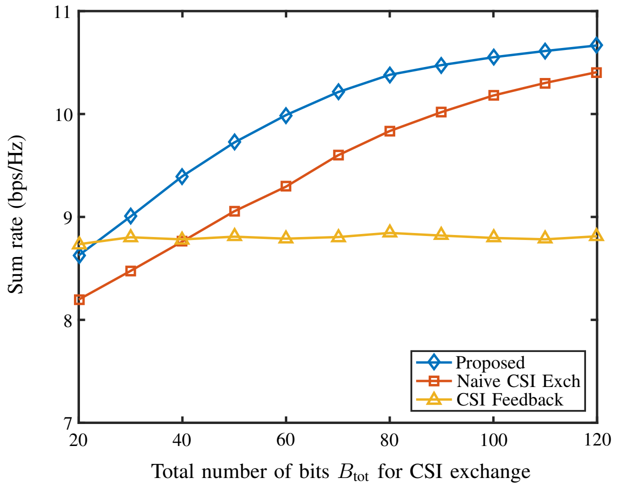

Figure 5. Sum rate versus total number of bits B_{d} for CSI exchange.

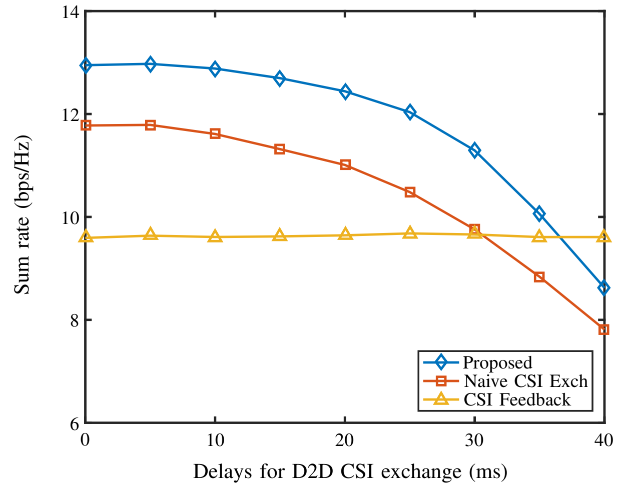

Figure 6. Sum rate versus various delays for CSI exchange over D2D.

at the BS is 20 dB. For less than 20 ms delay, the sum rate performance is only slightly affected, since the performance loss is dominated by the quantization error in D2D CSI exchange. In general, the proposed scheme can tolerate more than 30 ms delay for CSI exchange to perform better than the traditional CSI feedback scheme.

VI. Conclusions

We proposed a cooperative precoder feedback strategy for multiuser downlink transmission in FDD massive MIMO systems. The strategy consists of two phases. First, the users exploit reliable D2D communication to exchange the CSI, and second, the users individually compute the precoder and feed back the precoder to the BS. We analyzed the interference leakage when users have identically uncorrelated channel statistics. Our results showed that the precoder feedback scheme can reduce the interference leakage to $1/(K-1)$ of the CSI feedback scheme with ZF precoding. When users have non-identical channel statistics, we developed novel adaptive CSI exchange strategy, which exploits the global CSI statistics of the users. Optimal bit partition algorithm was derived for CSI exchange in terms of maximizing the virtual SLNR. Numerical results demonstrated that the proposed precoder feedback scheme with adaptive CSI exchange significantly outperforms the CSI feedback scheme in terms of higher throughput and lower feedback. The results also showed that the proposed scheme significantly saves the D2D overhead for CSI exchange.

Appendix A
Proof of Lemma 1

We first note that, vectors $\mathbf{s}_j$ follow an isotropic distribution in the (N_t-1)-dimensional subspace, because both of the quantization vectors $\hat{\mathbf{g}}_j$ from the RVQ codebook and the channel direction vectors $\mathbf{g}_j$ are isotropically distributed in the N_t-dimensional sphere. Thus, for any unit norm vector $\mathbf{w}$ independent of $\mathbf{s}_j$, $\mathbb{E}_{\mathcal{H},\mathcal{C}}\{\mathbf{s}_j^{\mathrm{H}}\mathbf{w}\,|\mathbf{w}\} = 0$. Therefore, the following holds from the property of iterative expectation

$$\begin{aligned}\mathbb{E}_{\mathcal{H},\mathcal{C}}&\left\{\mathbf{w}_1^{\mathrm{H}}\hat{\mathbf{g}}_j\mathbf{s}_j^{\mathrm{H}}\mathbf{w}_1\right\}\\ &=\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{\mathbf{w}_1^{\mathrm{H}}\hat{\mathbf{g}}_j\,\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{\mathbf{s}_j^{\mathrm{H}}\mathbf{w}_1\middle|\mathbf{w}_1,\hat{\mathbf{g}}_j\right\}\right\}\\ &=0. \end{aligned}\tag{30}$$

Similarly, $\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{\mathbf{w}_1^{\mathrm{H}}\mathbf{s}_j\hat{\mathbf{g}}_j^{\mathrm{H}}\mathbf{w}_1\right\}=0$. As $\mathbf{g}_j=\sqrt{1-Z_j}\hat{\mathbf{g}}_j+\sqrt{Z_j}\mathbf{s}_j$ from the definition of Z_j and $\mathbf{s}_j$, the following holds

$$\begin{aligned}&\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_1\right\}\\ &=\rho\sum_{j\neq 1}\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{\|\mathbf{h}_j\|^2|\mathbf{g}_j^{\mathrm{H}}\mathbf{w}_1|^2\right\}\\ &\overset{(a)}{=}\rho N_t\sum_{j\neq 1}\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{\left|\sqrt{1-Z_j}\hat{\mathbf{g}}_j^{\mathrm{H}}\mathbf{w}_1+\sqrt{Z_j}\mathbf{s}_j^{\mathrm{H}}\mathbf{w}_1\right|^2\right\}\\ &=\rho N_t\sum_{j\neq 1}\mathbb{E}_{\mathcal{H},\mathcal{C}}\Big\{(1-Z_j)\left|\hat{\mathbf{g}}_j^{\mathrm{H}}\mathbf{w}_1\right|^2+Z_j\left|\mathbf{s}_j^{\mathrm{H}}\mathbf{w}_1\right|^2\\ &\qquad+\sqrt{(1-Z_j)Z_j}\left[\mathbf{w}_1^{\mathrm{H}}\hat{\mathbf{g}}_j\mathbf{s}_j^{\mathrm{H}}\mathbf{w}_1+\mathbf{w}_1^{\mathrm{H}}\mathbf{s}_j\hat{\mathbf{g}}_j^{\mathrm{H}}\mathbf{w}_1\right]\Big\}\\ &\overset{(b)}{=}\rho N_t\sum_{j\neq 1}\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{(1-Z_j)\left|\hat{\mathbf{g}}_j^{\mathrm{H}}\mathbf{w}_1\right|^2+Z_j\left|\mathbf{s}_j^{\mathrm{H}}\mathbf{w}_1\right|^2\right\}.\end{aligned}$$

where $\overset{(a)}{=}$ is due to the fact that the channel magnitude $\|\mathbf{h}_j\|^2$ is independent of both $\mathbf{g}_j$ (channel direction) and $\mathbf{w}_1$ (precoding based on $\{\mathbf{g}_j\}$), and $\overset{(b)}{=}$ is due to (30).

Appendix B
Proof of Theorem 2

Lemma 7 (Interference due to Discrete Precoding): The random variable $\left|\hat{\mathbf{g}}_2^{\mathrm{H}}\mathbf{w}_1^{\mathrm{c}}\right|^2$ follows a beta distribution $\mathcal{B}(1,(N_t-1)2^{B_{\mathrm{f}}})$ and its mean is given by $\left(1+(N_t-1)2^{B_{\mathrm{f}}}\right)^{-1}$.

Proof: By the construction of a RVQ codebook $\mathcal{C}_1^{\mathrm{w}}$, the codewords $\mathbf{w}\in\mathcal{C}_1^{\mathrm{w}}$ follow isotropic distribution in the N_t-dimensional subspace. Thus $\left|\hat{\mathbf{g}}_2^{\mathrm{H}}\mathbf{w}^{\mathrm{c}}\right|^2$ follows $\mathcal{B}(1,N_t-1)$

distribution, with CDF given by

$$\mathbb{P}\left\{\left|\hat{\mathbf{g}}_2^{\mathrm{H}}\mathbf{w}^{\mathrm{c}}\right|^2 \le x\right\} = 1-(1-x)^{N_t-1}.$$

As a result,

$$\begin{aligned}\mathbb{P}\left\{\left|\hat{\mathbf{g}}_2^{\mathrm{H}}\mathbf{w}_1^{\mathrm{c}}\right|^2 \le x\right\} &= \mathbb{P}\left\{\min_{\mathbf{w}^{\mathrm{c}}\in\mathcal{C}_1^{\mathrm{w}}}\left|\hat{\mathbf{g}}_2^{\mathrm{H}}\mathbf{w}^{\mathrm{c}}\right|^2 \le x\right\}\\ &= 1-\mathbb{P}\left\{\left|\hat{\mathbf{g}}_2^{\mathrm{H}}\mathbf{w}_i^{\mathrm{c}}\right|^2 > x,\, \forall \mathbf{w}_i^{\mathrm{c}}\in\mathcal{C}_1^{\mathrm{w}}\right\}\\ &= 1-\mathbb{P}\left\{\left|\hat{\mathbf{g}}_2^{\mathrm{H}}\mathbf{w}^{\mathrm{c}}\right|^2 > x\right\}^{N_{\mathrm{f}}}\\ &= 1-(1-x)^{(N_t-1)N_{\mathrm{f}}}\end{aligned}$$

which means that $\min_{\mathbf{w}\in\mathcal{C}_1^{\mathrm{w}}}\left|\hat{\mathbf{g}}_2^{\mathrm{H}}\mathbf{w}\right|^2$ follows the beta distribution $\mathcal{B}(1,(N_t-1)N_{\mathrm{f}})$, where $N_{\mathrm{f}}=|\mathcal{C}_1^{\mathrm{w}}|=2^{B_{\mathrm{f}}}$.

Moreover, the mean of a beta random variable $\mathcal{B}(\alpha,\beta)$ is given by $\alpha/(\alpha+\beta)$, and hence $\mathbb{E}\left\{\left|\hat{\mathbf{g}}_2^{\mathrm{H}}\mathbf{w}_1^{\mathrm{c}}\right|^2\right\} = 1/[1+(N_t-1)2^{B_{\mathrm{f}}}]$. □

From Lemma 1 and 7 and the channel quantization error bounds in (6), we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{I_1^{\mathrm{c}}\right\}\\ &= \rho N_t \mathbb{E}_{\mathcal{H},\mathcal{C}}\left\{(1-Z_2)\left|\hat{\mathbf{g}}_2^{\mathrm{H}}\mathbf{w}_1^{\mathrm{c}}\right|^2 + Z_2\left|\mathbf{s}_2^{\mathrm{H}}\mathbf{w}_1^{\mathrm{c}}\right|^2\right\}\\ &\le \rho N_t\left[\left(1-\frac{N_t-1}{N_t}2^{-\frac{B_{\mathrm{c}}}{N_t-1}}\right)\frac{1}{1+(N_t-1)2^{B_{\mathrm{f}}}}\right.\\ &\qquad\qquad\left.+\,2^{-\frac{B_{\mathrm{c}}}{N_t-1}}\times\frac{1}{N_t-1}\right]\\ &\le \rho N_t\left[\left(1-\frac{N_t-1}{N_t}2^{-\frac{B_{\mathrm{c}}}{N_t-1}}\right)\frac{2^{-B_{\mathrm{f}}}}{N_t-1}+\frac{2^{-B_{\mathrm{c}}/(N_t-1)}}{N_t-1}\right]\\ &= \frac{\rho N_t}{N_t-1}\left[2^{-B_{\mathrm{f}}}+\left(1-\frac{N_t-1}{N_t}2^{-B_{\mathrm{f}}}\right)2^{-\frac{B_{\mathrm{c}}}{N_t-1}}\right].\end{aligned}$$

Appendix C
Proof of Lemma 2

We first show that $Y_{j,i} \triangleq |\hat{\mathbf{h}}_j^{\mathrm{H}}\tilde{\mathbf{w}}_i|^2 \xrightarrow{\mathrm{d}} \frac{1}{2}\chi^2(2)$, as $N_t\to\infty$, where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution. Note that

$$Y_{j,i} = \|\hat{\mathbf{h}}_j\|^2|\hat{\mathbf{u}}_j^{\mathrm{H}}\tilde{\mathbf{w}}_i|^2 = \frac{1}{N_t}\|\hat{\mathbf{h}}_j\|^2\cdot N_t Z$$

where $\hat{\mathbf{u}}_j = \hat{\mathbf{h}}_j/\|\hat{\mathbf{h}}_j\|$ and $Z = |\hat{\mathbf{u}}_j^{\mathrm{H}}\tilde{\mathbf{w}}_i|^2$ is well-known to follow the beta distribution $\mathcal{B}(1,N_t-1)$, because $\hat{\mathbf{u}}_j$ is a unit norm N_t-dimensional vector and $\mathbf{w}_i$ is random, isotropic, and independent to $\hat{\mathbf{u}}_j$.

Note that each element of $\mathbf{h}_j$ is i.i.d. Gaussian $\mathcal{CN}(0,1)$. Then, by the Strong Law of Large Numbers,

$$\begin{aligned}\frac{1}{N_t}\|\hat{\mathbf{h}}_j\|^2 &= \frac{1}{N_t}\|\mathbf{h}_j\|^2\\ &= \frac{1}{N_t}\left(|h_{j1}|^2+|h_{j2}|^2+\cdots+|h_{jN_t}|^2\right)\xrightarrow{\mathrm{a.s.}}1\end{aligned}\tag{31}$$

as $N_t\to\infty$, where $\xrightarrow{\mathrm{a.s.}}$ denotes almost surely (a.s.) convergence.

Let Z be a beta random variable following $\mathcal{B}(1,N_t-1)$ distribution. Then

$$\begin{aligned}\lim_{N_t\to\infty}\mathbb{P}\left\{N_t Z\le x\right\} &= \lim_{N_t\to\infty}\mathbb{P}\left\{Z\le\frac{x}{N_t}\right\}\\ &= \lim_{N_t\to\infty}1-(1-\frac{x}{N_t})^{N_t-1}\\ &= \lim_{N_t\to\infty}1-(1-\frac{x}{N_t})^{N_t}\frac{1}{1-x/N_t}\\ &= 1-e^{-x}\end{aligned}$$

On the other hand, $\mathbb{P}\left\{\frac{1}{2}\chi^2(2)\le x\right\} = 1-e^{-x}$, which shows that $N_t Z\xrightarrow{\mathrm{d}}\frac{1}{2}\chi^2(2)$. Using (31), we can conclude that $Y_{j,i}\xrightarrow{\mathrm{d}}\frac{1}{2}\chi^2(2)$.

We then show that $Y_{j,i}$'s are mutually independent with respect to (w.r.t.) j. First, from the independency of $\mathbf{h}_j$, the quantized vectors $\hat{\mathbf{h}}_j$ are independent. In addition, from $Y_{j,i} = \|\hat{\mathbf{h}}_j\|^2|\hat{\mathbf{u}}_j^{\mathrm{H}}\tilde{\mathbf{w}}_i|^2$, the random variables $|\hat{\mathbf{u}}_j^{\mathrm{H}}\tilde{\mathbf{w}}_i|^2$ and $|\hat{\mathbf{u}}_k^{\mathrm{H}}\tilde{\mathbf{w}}_i|^2$, $k\ne j$, are mutually independent, because $\hat{\mathbf{u}}_j$ are independently and isotropically distributed. These conclude that $Y_{j,i}$'s are mutually independent w.r.t. j.

As a result, $\widetilde{Y}_i = \sum_{j\ne 1}Y_{j,i}$ converges to the sum of $K-1$ i.i.d. $\frac{1}{2}\chi^2(2)$ random variables, which is $\frac{1}{2}\chi^2(2(K-1))$.

In addition, given $\{\hat{\mathbf{h}}_j\}_{j\ne1}$, $\widetilde{Y}_i$ and $\widetilde{Y}_l$ are independent. The independence of $|\hat{\mathbf{u}}_j^{\mathrm{H}}\tilde{\mathbf{w}}_i|^2$ and $|\hat{\mathbf{u}}_j^{\mathrm{H}}\tilde{\mathbf{w}}_l|^2$ follows from the independence between isotropic random vectors $\mathbf{w}_i$. As $\frac{1}{N_t}\|\hat{\mathbf{h}}_j\|^2\xrightarrow{\mathrm{a.s.}}1$, $\widetilde{Y}_i$ and $\widetilde{Y}_l$ become asymptotically independent for large N_t. Hence, $(\widetilde{Y}_1,\widetilde{Y}_2,\ldots,\widetilde{Y}_N)$ converges to $\frac{1}{2}(X_1,X_2,\ldots,X_N)$ in distribution.

Appendix D
Proof of Lemma 3

From Lemma 2, as $N_t\to\infty$, $\widetilde{Y}_i = \sum_{j\ne k}|\hat{\mathbf{h}}_j^{\mathrm{H}}\mathbf{w}_i|^2$ converges to i.i.d. chi-square random variables $\frac{1}{2}\chi^2(2(K-1))$. The limiting CDF of $\widetilde{Y}_i$ is thus given by $F_K(y) = \frac{1}{\Gamma(K-1)}\overline{\gamma}(y;K-1)$, $y\ge0$, where $\overline{\gamma}(y;k)=\int_0^y u^{k-1}e^{-u}du$ is the incomplete gamma function.

Define $F_K^*(y) = F_K(-\frac{1}{y})$ for $y\le0$. Then the following property holds

$$\begin{aligned}\lim_{t\to-\infty}\frac{F_K^*(ty)}{F_K^*(t)} &= \lim_{t\to-\infty}\frac{\overline{\gamma}(-\frac{1}{2ty};K-1)}{\overline{\gamma}(-\frac{1}{2t};K-1)}\\ &= \lim_{t\to-\infty}\frac{\overline{\gamma}(-\frac{1}{2ty};K-1)}{\left(-\frac{1}{2ty}\right)^{K-1}}\frac{\left(-\frac{1}{2t}\right)^{K-1}}{\overline{\gamma}(-\frac{1}{2t};K-1)}\frac{\left(-\frac{1}{2ty}\right)^{K-1}}{\left(-\frac{1}{2t}\right)^{K-1}}\\ &= y^{-(K-1)}\end{aligned}\tag{32}$$

where we used the property of incomplete gamma function that $\lim_{x\to0}\overline{\gamma}(x;k)/x^k = \frac{1}{k}$.

The extreme value theory [41, Theorem 2.1.5] concludes that under condition (32),

$$\lim_{N\to\infty}\lim_{N_t\to\infty}\mathbb{P}\left\{\min_{i=1,2,\ldots,N}\widetilde{Y}_i<\phi_N y\right\} = 1-\exp(-y^{K-1})$$

for $y\ge0$, where $\phi_N=\sup\{y: F_K(y)\le\frac{1}{N}\}$ which yields (8).

Moreover, using the limiting property of the incomplete gamma function $\overline{\gamma}(y;k) = \frac{y^k}{k} + o(y^k)$, we have $F_K(y) \approx \frac{y^{K-1}}{(K-1)\Gamma(K-1)} = \frac{y^{K-1}}{\Gamma(K)}$. Solving

$$F_K(y) \approx \frac{y^{K-1}}{\Gamma(K)} = \frac{1}{N}$$

for y, gives $\phi_N \approx \hat{y} = \Gamma(K)^{-\frac{1}{K-1}} N^{-\frac{1}{K-1}}$ as in (9).

Note that since the $F_K(y)$ decreases when K increases, thus the optimal solution $y^* = F_K^{-1}(\frac{1}{N})$ decreases as K decreases. Meanwhile, the approximation $F_K(y) \approx y^{K-1}/\Gamma(K)$ is asymptotically accurate when y approaches 0. This means that the approximation of ϕ_N becomes accurate for small K.

Appendix E
Proof of Lemma 5

Under the KLT (20), the vector $\mathbf{q}_k^{(j)}$ has independent elements, where the ith complex element $\left[\mathbf{q}_k^{(j)}\right]_i$ has variance $\lambda_{kj}^{(i)}$. According to Shannon's distortion-rate theory, the minimum distortion of the ith complex element is given by

$$D_{kj}^{(i)} \triangleq \mathbb{E}\left\{\left(\left[\mathbf{q}_k^{(j)}\right]_i - \left[\hat{\mathbf{q}}_k^{(j)}\right]_i\right)^2\right\} = \lambda_{kj}^{(i)} 2^{-b_{kj}^{(i)}}$$

where $b_{kj}^{(i)}$ is the number of bits allocated to the ith complex element of $\mathbf{q}_k^{(j)}$. Therefore, the minimum distortion $D_{kj} = \sum_{i=1}^{M_{kj}} D_{kj}^{(i)}$ can be achieved by

$$\begin{aligned} \underset{\{b_{kj}^{(i)} \geq 0\}}{\text{minimize}} \quad & D_{kj} = \sum_{i=1}^{M_{kj}} \lambda_{kj}^{(i)} 2^{-b_{kj}^{(i)}} \\ \text{subject to} \quad & \sum_{i=1}^{M_{kj}} b_{kj}^{(i)} = b_{kj} \end{aligned} \tag{33}$$

Lemma 8: With sufficiently large b_{kj}, the minimum value of (33) is given by

$$D_{kj}^* = M_{kj} \Big(\prod_{i=1}^{M_{kj}} \lambda_{kj}^{(i)} \Big)^{\frac{1}{M_{kj}}} 2^{-\frac{b_{kj}}{M_{kj}}}$$

and the distortion of each element $\left[\mathbf{q}_k^{(j)}\right]_i$ is $D_{kj}^{(i)} = \frac{1}{M_{kj}} D_{kj}^*$, for $i = 1, 2, \ldots, M_{kj}$.

Proof: The closed-form solution to (33) can be derived using Lagrangian methods. Details are omitted here due to page limit. □

As a result of Lemma 8, we have $\mathbf{q}_k^{(j)} - \hat{\mathbf{q}}_k^{(j)} \sim \mathcal{CN}(\mathbf{0}, \frac{D_{kj}^*}{M_{kj}} \mathbf{I})$. From (20) and (18), we have $\mathbf{h}_k^{(j)} = \sqrt{l_k} \mathbf{U}_j \mathbf{U}_{kj} \mathbf{q}_k^{(j)}$, and hence,

$$\mathbf{h}_k^{(j)} - \hat{\mathbf{h}}_k^{(j)} \sim \mathcal{CN}(\mathbf{0}, l_k \frac{D_{kj}^*}{M_{kj}} \mathbf{U}_j \mathbf{U}_{kj} \mathbf{U}_{kj}^{\mathrm{H}} \mathbf{U}_j^{\mathrm{H}}).$$

In addition, from the CSI decomposition model (19),

$$\mathbf{h}_k^{(j)\perp} \sim \mathcal{CN}(\mathbf{0}, (\mathbf{I} - \mathbf{U}_j \mathbf{U}_j^{\mathrm{H}}) \mathbf{R}_k (\mathbf{I} - \mathbf{U}_j \mathbf{U}_j^{\mathrm{H}})^{\mathrm{H}})$$

which is orthogonal to $\mathbf{h}_k^{(j)}$ and is not exchanged between users. Therefore, $\mathbf{h}_k - \hat{\mathbf{h}}_k^{(j)}$ is zero mean with covariance given by

$$l_k \frac{D_{kj}^*}{M_{kj}} \mathbf{U}_j \mathbf{U}_{kj} \mathbf{U}_{kj}^{\mathrm{H}} \mathbf{U}_j^{\mathrm{H}} + (\mathbf{I} - \mathbf{U}_j \mathbf{U}_j^{\mathrm{H}}) \mathbf{R}_k (\mathbf{I} - \mathbf{U}_j \mathbf{U}_j^{\mathrm{H}})^{\mathrm{H}}$$

which confirms the result in Lemma 5.

Appendix F
Proof of Lemma 6

The virtual SLNR $\bar{\Gamma}_k$ can be lower bounded by

$$\begin{aligned} \bar{\Gamma}_k &= \mathbb{E}\left\{\frac{|\mathbf{h}_k^{\mathrm{H}} \tilde{\mathbf{w}}_k^{\mathrm{c}}|^2}{\sum_{j \neq k} |\mathbf{h}_j^{\mathrm{H}} \tilde{\mathbf{w}}_k^{\mathrm{c}}|^2 + \alpha}\right\} \\ &= \mathbb{E}\left\{\mathbb{E}\left\{\frac{|\mathbf{h}_k^{\mathrm{H}} \tilde{\mathbf{w}}_k^{\mathrm{c}}|^2}{\sum_{j \neq k} |\mathbf{h}_j^{\mathrm{H}} \tilde{\mathbf{w}}_k^{\mathrm{c}}|^2 + \alpha} \middle| \hat{\mathbf{H}}_k\right\}\right\} \\ &\overset{(a)}{\geq} \mathbb{E}\left\{\frac{|\mathbf{h}_k^{\mathrm{H}} \tilde{\mathbf{w}}_k^{\mathrm{c}}|^2}{\sum_{j \neq k} |\mathbf{h}_j^{\mathrm{H}} \tilde{\mathbf{w}}_k^{\mathrm{c}}|^2 + (\tilde{\mathbf{w}}_k^{\mathrm{c}})^{\mathrm{H}} \sum_{j \neq k} \mathbf{Q}_{jk} \tilde{\mathbf{w}}_k^{\mathrm{c}} + \alpha}\right\} \\ &\overset{(b)}{\geq} \mathbb{E}\left\{\frac{|\mathbf{h}_k^{\mathrm{H}} \mathbf{w}_k^{\mathrm{ZF}}|^2}{(\mathbf{w}_k^{\mathrm{ZF}})^{\mathrm{H}} \sum_{j \neq k} \mathbf{Q}_{jk} \mathbf{w}_k^{\mathrm{ZF}} + \alpha}\right\} \end{aligned} \tag{34}$$

where $\tilde{\mathbf{w}}_k^{\mathrm{c}}$ is the SLNR precoder in continuous domain obtained from (15) by replacing $\mathbf{w} \in \mathcal{C}_k^{\mathrm{w}}$ to $\|\mathbf{w}\| = 1$, $\mathbf{w}_k^{\mathrm{ZF}} = \tilde{\mathbf{w}}_k^{\mathrm{ZF}} / \|\tilde{\mathbf{w}}_k^{\mathrm{ZF}}\|$ and $\tilde{\mathbf{w}}_k^{\mathrm{ZF}}$ is the kth column of the ZF precoding matrix $\tilde{\mathbf{W}}_k = \hat{\mathbf{H}}_k (\hat{\mathbf{H}}_k^{\mathrm{H}} \hat{\mathbf{H}}_k)^{-1}$. Inequality $\overset{(a)}{\geq}$ is from (13), and $\overset{(b)}{\geq}$ is due to the fact that $\mathbf{w}_k^{\mathrm{ZF}}$ is not optimal in maximizing the SLNR criterion (15).

A. The Interference Term

Consider the result in Lemma 5. We know that

$$(\mathbf{w}_k^{\mathrm{ZF}})^{\mathrm{H}} (\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^{\mathrm{H}}) \mathbf{R}_j (\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^{\mathrm{H}})^{\mathrm{H}} \mathbf{w}_k^{\mathrm{ZF}} = 0$$

since by construction (from a projection of $\mathbf{h}_k$), $\mathbf{w}_k^{\mathrm{ZF}}$ lies in the subspace spanned by the columns of $\mathbf{U}_k$. In addition,

$$(\mathbf{w}_k^{\mathrm{ZF}})^{\mathrm{H}} \mathbf{U}_k \mathbf{U}_{jk} \mathbf{U}_{jk}^{\mathrm{H}} \mathbf{U}_k^{\mathrm{H}} \mathbf{w}_k^{\mathrm{ZF}} \leq 1$$

since none of the the eigenvalues of the matrix $\mathbf{U}_k \mathbf{U}_{jk} \mathbf{U}_{jk}^{\mathrm{H}} \mathbf{U}_k^{\mathrm{H}}$ is larger than 1.

As a result of Lemma 5, we have

$$(\mathbf{w}_k^{\mathrm{ZF}})^{\mathrm{H}} \mathbf{Q}_{jk} \mathbf{w}_k^{\mathrm{ZF}} \leq l_j \Big(\prod_{m=1}^{M_{jk}} \lambda_{jk}^{(m)} \Big)^{\frac{1}{M_{jk}}} 2^{-\frac{b_{jk}}{M_{jk}}}.$$

B. The signal term

The signal term $\mathbb{E}\left\{|\mathbf{h}_k^{\mathrm{H}} \mathbf{w}_k^{\mathrm{ZF}}|^2\right\}$ can be computed as follows. Let

$$\mathbf{P}_k = \mathbf{I} - \hat{\mathbf{H}}_{-k} \big(\hat{\mathbf{H}}_{-k}^{\mathrm{H}} \hat{\mathbf{H}}_{-k} \big)^{-1} \hat{\mathbf{H}}_{-k}^{\mathrm{H}}$$

be a $N_t \times N_t$ projection matrix for user k, where $\hat{\mathbf{H}}_{-k} = \left[\left\{\hat{\mathbf{h}}_j^{(k)} : j \neq k\right\}\right]$ is a $N_t \times (K-1)$ CSI matrix that contains the CSI exchanged from all the other users. As a result, the ZF precoder $\mathbf{w}_k^{\mathrm{ZF}}$ can be equivalently written as

$$\mathbf{w}_k^{\mathrm{ZF}} = \frac{\mathbf{P}_k \mathbf{h}_k}{\|\mathbf{P}_k \mathbf{h}_k\|}.$$

Using the property of a projection matrix $\mathbf{P}_k = \mathbf{P}_k \mathbf{P}_k^{\mathrm{H}} = \mathbf{P}_k^{\mathrm{H}}$, the following holds

$$|\mathbf{h}_k^{\mathrm{H}} \mathbf{w}_k^{\mathrm{ZF}}|^2 = \frac{|\mathbf{h}_k^{\mathrm{H}} \mathbf{P}_k \mathbf{h}_k|^2}{\|\mathbf{P}_k \mathbf{h}_k\|^2} = \frac{\|\mathbf{h}_k^{\mathrm{H}} \mathbf{P}_k^{\mathrm{H}} \mathbf{P}_k \mathbf{h}_k\|^2}{\|\mathbf{P}_k \mathbf{h}_k\|^2} = \|\mathbf{P}_k \mathbf{h}_k\|^2.$$

As a result,

$$\begin{aligned}\mathbb{E}\left\{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_k^{\mathrm{ZF}}|^2\middle|\hat{\mathbf{H}}_{-k}\right\} &= \mathbb{E}\left\{\|\mathbf{P}_k\mathbf{h}_k\|^2\middle|\hat{\mathbf{H}}_{-k}\right\}\\ &= \mathbb{E}\left\{\mathrm{tr}\{\mathbf{P}_k\mathbf{h}_k\mathbf{h}_k^{\mathrm{H}}\mathbf{P}_k^{\mathrm{H}}\}\middle|\hat{\mathbf{H}}_{-k}\right\}\\ &\overset{(a)}{=} \mathrm{tr}\left\{\mathbf{P}_k\mathbb{E}\left\{\mathbf{h}_k\mathbf{h}_k^{\mathrm{H}}\middle|\hat{\mathbf{H}}_{-k}\right\}\mathbf{P}_k^{\mathrm{H}}\right\}\\ &\overset{(b)}{=} \mathrm{tr}\left\{\mathbf{P}_k l_k\mathbf{R}_k\mathbf{P}_k^{\mathrm{H}}\right\}\\ &\overset{(c)}{\geq} l_k\sum_{i=K}^{N_t}\lambda_k^{(i)}\end{aligned}$$

where $\lambda_k^{(i)}$ are the eigenvalues of $\mathbf{R}_k$ in descending order, the equality $\overset{(a)}{=}$ is because $\mathbf{P}_k$ only depends on $\hat{\mathbf{H}}_{-k}$, the equality $\overset{(b)}{=}$ is due to the independence between $\mathbf{h}_k$ and $\hat{\mathbf{H}}_{-k}$, and the lower bound $\overset{(c)}{\geq}$ is tight when $\mathbf{P}_k$ is to project $\mathbf{R}_k$ onto the orthogonal subspace of the subspace that is spanned by the $K-1$ dominant eigenvectors of $\mathbf{R}_k$.

Therefore,

$$\mathbb{E}\left\{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_k^{\mathrm{ZF}}|^2\right\} = \mathbb{E}\left\{\mathbb{E}\left\{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_k^{\mathrm{ZF}}|^2\middle|\hat{\mathbf{H}}_{-k}\right\}\right\} \geq l_k\sum_{i=K}^{N_t}\lambda_k^{(i)}.$$

C. The Lower Bound

The virtual SLNR can be further bounded as

$$\begin{aligned}\bar{\Gamma}_k &\geq \mathbb{E}\left\{\frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_k^{\mathrm{ZF}}|^2}{(\mathbf{w}_k^{\mathrm{ZF}})^{\mathrm{H}}\sum_{j\neq k}\mathbf{Q}_{jk}\mathbf{w}_k^{\mathrm{ZF}}+\alpha}\right\}\\ &\geq \frac{l_k\sum_{i=K}^{N_t}\lambda_k^{(i)}}{\sum_{j\neq k}l_j\left(\prod_{m=1}^{M_{jk}}\lambda_{jk}^{(m)}\right)^{\frac{1}{M_{jk}}}2^{-\frac{b_{jk}}{M_{jk}}}+\alpha}\end{aligned}$$

which proves the result.

Appendix G
Proof of Theorem 4

As the constrained minimization problem (27) is convex, it can be solved using Lagrangian methods. Specifically, the Lagrangian function of (27) can be written as

$$\mathcal{L}(\mathbf{b},\mu) = \sum_{k=1}^{K}\log\left(\sum_{j\neq k}\omega_{jk}2^{-\frac{b_{jk}}{M_{jk}}}+\alpha\right)+\mu\left[\sum_{k=1}^{K}\sum_{j\neq k}b_{kj}-B_{\mathrm{tot}}\right]$$

and the Karush-Kuhn-Tucker (KKT) condition is given by

$$\frac{\partial\mathcal{L}(\mathbf{b},\mu)}{\partial b_{jk}} = \frac{-\frac{\ln 2}{M_{jk}}\omega_{jk}2^{-\frac{b_{jk}}{M_{jk}}}}{\sum_{m\neq k}\omega_{mk}2^{-\frac{b_{mk}}{M_{mk}}}+\alpha}+\mu = 0,\ b_{jk}\geq 0 \qquad (35)$$

$$\forall j\neq k,\ k=1,2,\ldots,K$$

$$\mu\left(\sum_{k=1}^{K}\sum_{j\neq k}b_{kj}-B_{\mathrm{tot}}\right) = 0, \qquad \mu\geq 0. \qquad (36)$$

Condition (35) can be divided into K sets of equations. Each set consists of $K-1$ equations as follows

$$\frac{\ln 2}{M_{jk}}\omega_{jk}2^{-\frac{b_{jk}}{M_{jk}}} - \mu\sum_{m\neq k}\omega_{mk}2^{-\frac{b_{mk}}{M_{mk}}} = \mu\alpha, \qquad j\neq k$$

which can be written into a compact form as

$$\mathbf{A}_k\mathbf{x}_k - \mu\mathbf{1}\boldsymbol{\omega}_k^{\mathrm{T}}\mathbf{x}_k = \mu\alpha$$

where

$$\mathbf{x}_k = (2^{-\frac{b_{1k}}{M_{1k}}},2^{-\frac{b_{2k}}{M_{2k}}},\ldots,2^{-\frac{b_{k-1,k}}{M_{k-1,k}}},2^{-\frac{b_{k+1,k}}{M_{k+1,k}}},\ldots,2^{-\frac{b_{Kk}}{M_{Kk}}})^{\mathrm{T}}.$$

This leads to solutions (28) and (29), where the projection $[\cdot]^+$ and the choice of μ are to satisfy the KKT conditions (28) and (36).

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, 2013.

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

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

[4] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving "massive MIMO" spectral efficiency with a not-so-large number of antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3226–3239, 2012.

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

[6] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, 2011.

[7] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 264–273, 2013.

[8] 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. Areas Commun.*, vol. 8, no. 5, pp. 802–814, 2014.

[9] S. Noh, M. D. Zoltowski, and D. J. Love, "Training sequence design for feedback assisted hybrid beamforming in massive MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 187–200, 2016.

[10] D. J. Love and R. W. Heath, "Limited feedback diversity techniques for correlated channels," *IEEE Trans. Veh. Technol.*, vol. 55, no. 2, pp. 718–722, 2006.

[11] V. Raghavan, R. W. Heath, and A. M. Sayeed, "Systematic codebook designs for quantized beamforming in correlated MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1298–1310, 2007.

[12] D. Ying, F. W. Vook, T. A. Thomas, D. J. Love, and A. Ghosh, "Kronecker product correlation model and limited feedback codebook design in a 3D channel model," in *Proc. IEEE Int. Conf. Commun.*, 2014, pp. 5865–5870.

[13] J. Choi, B. Clerckx, N. Lee, and G. Kim, "A new design of polar-cap differential codebook for temporally/spatially correlated MISO channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 703–711, 2012.

[14] V. Raghavan, V. V. Veeravalli, and A. M. Sayeed, "Quantized multimode precoding in spatially correlated multiantenna channels," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 6017–6030, 2008.

[15] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5060, 2006.

[16] J. Choi, Z. Chance, D. J. Love, and U. Madhow, "Noncoherent trellis coded quantization: A practical limited feedback technique for massive MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 12, pp. 5016–5029, 2013.

[17] J. Choi, D. J. Love, and T. Kim, "Trellis-extended codebooks and successive phase adjustment: A path from LTE-Advanced to FDD massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2007–2016, 2015.

[18] X. Rao and V. K. 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, 2014.

[19] 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, 2015.

[20] C. Hao, Y. Wu, and B. Clerckx, "Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3232–3246, 2015.

[21] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Trans. Wireless Commun.*, no. 7, pp. 4611–4624, 2016.

[22] A. Adhikary, J. Nam, J. Ahn, and G. Caire, "Joint spatial division and multiplexing – The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441 – 6463, Oct 2013.

[23] J. Nam, A. Adhikary, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 876–890, Oct 2014.

[24] J. Chen and V. K. N. Lau, "Two-tier precoding for FDD multi-cell massive MIMO time-varying interference networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1230–1238, Jun. 2014.

[25] D. Kim, G. Lee, and Y. Sung, "Two-stage beamformer design for massive MIMO downlink by trace quotient formulation," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2200–2211, Jun. 2015.

[26] H. Kwon and J. M. Cioffi, "MISO broadcast channel with user-cooperation and limited feedback," in *Proc. IEEE Int. Symp. on Inf. Theory*, 2009, pp. 1694–1698.

[27] ——, "Multi-user MISO broadcast channel with user-cooperating decoder," in *Proc. IEEE Semiannual Veh. Technol. Conf.*, 2008, pp. 1–5.

[28] H. Mao, W. Feng, and N. Ge, "Receiver cooperation for MIMO broadcast channels with finite-rate feedback," *IEEE Commun. Lett.*, vol. 19, no. 5, pp. 887–890, 2015.

[29] P.-H. Kuo, J. Choi, J. Suh, and S. Kim, "A feedback scheme for ZFBF-based MIMO broadcast systems with infrastructure relay stations," in *Proc. IEEE Wireless Commun. and Networking Conf.*, 2008, pp. 1061–1066.

[30] D. J. Love, R. W. Heath, V. K. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, 2008.

[31] H. Yin, L. Cottatellucci, and D. Gesbert, "Enabling massive MIMO systems in the FDD mode thanks to D2D communications," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, Nov. 2014, pp. 656–660.

[32] J. Chen, H. Yin, L. Cottatellucci, and D. Gesbert, "Precoder feedback versus channel feedback in massive MIMO under user cooperation," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, Nov. 2015, pp. 1449–1453.

[33] C. B. Peel, B. M. Hochwald *et al.*, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: channel inversion and regularization," *IEEE Trans. Wireless Commun.*, vol. 53, no. 1, pp. 195–202, 2005.

[34] M. Joham, W. Utschick, J. Nossek *et al.*, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, 2005.

[35] P. Patcharamaneepakorn, S. Armour, and A. Doufexi, "On the equivalence between SLNR and MMSE precoding schemes with single-antenna receivers," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 1034–1037, 2012.

[36] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Springer Science & Business Media, 1992, vol. 159.

[37] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1993.

[38] D. Marco and D. L. Neuhoff, "Low-resolution scalar quantization for Gaussian sources and squared error," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1689–1697, 2006.

[39] A. Forenza, D. J. Love, and R. W. Heath, "Simplified spatial correlation models for clustered MIMO channels with different array configurations," *IEEE Trans. Veh. Technol.*, vol. 56, no. 4, pp. 1924–1934, 2007.

[40] K. Baddour and N. Beaulieu, "Autoregressive models for fading channel simulation," in *Proc. IEEE Global Telecomm. Conf.*, vol. 2, 2001, pp. 1187 –1192 vol.2.

[41] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*, 2nd ed. Robert E. Krieger Pub. Co., Inc., 1987.

Junting Chen (S'11–M'16) received the Ph.D. degree in electronic and computer engineering from Hong Kong University of Science and Technology (HKUST), Hong Kong SAR China, in 2015, and the B.Sc. degree in electronic engineering from Nanjing University, Nanjing, China, in 2009.

He is a Postdoctoral Research Associate with the Department of Electrical Engineering, University of Southern California (USC), Los Angeles, CA, USA. Prior to that, he was a postdoctoral researcher with Department of Communication Systems at EURECOM, France, in 2015–2016. From 2014–2015, he was a visiting student with the Wireless Information and Network Sciences Laboratory at MIT, Cambridge, MA, USA. He was an recipient of the HKTIIT Post-Graduate Excellence Scholarships in 2012 from HKUST. His research interests include signal processing, optimizations, nonlinear control, and statistical learning, with applications to wireless communications and localization.

Haifan Yin obtained the Ph.D. degree from Télécom ParisTech, France in 2015. He received the B.Sc. degree and the M.Sc. degree from Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2012 respectively. He is currently a research engineer in Sequans Communications, France. His research interests include massive MIMO, signal processing, and cooperative networks. H. Yin was a recipient of the 2015 Chinese Government Award for Outstanding Self-financed Students Abroad.

Laura Cottatellucci (S'01–M'07) obtained the Ph.D. from Technical University of Vienna, Austria (2006) and the Master degree from La Sapienza University, Rome, Italy (1995). Specialized in networking at Guglielmo Reiss Romoli School (1996, Italy), she worked in Telecom Italia (1995–2000) as responsible of industrial projects. From April 2000 to September 2005 she worked as senior research in ftw Austria on CDMA and MIMO systems. She was research fellow in INRIA (Sophia Antipolis, France) from October to December 2005 and at the University of South Australia, Australia in 2006. Since December 2006 she is assistant professor in EURECOM. Cottatellucci is currently associate editor for IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON SIGNAL PROCESSING and served as guest editor for EURASIP JOURNAL ON WIRELESS COMMUNICATIONS AND NETWORKING (special issue on cooperative communications). Her research interests lie in the areas of large system analysis and algorithm design for multiuser wireless communications and complex networks, random matrix theory, and game theory.

David Gesbert (F'11) is Professor and Head of the Communication Systems Department, EURECOM. He obtained the Ph.D degree from Ecole Nationale Superieure des Telecommunications, France, in 1997. From 1997 to 1999 he has been with the Information Systems Laboratory, Stanford University. He was then a founding engineer of Iospan Wireless Inc, a Stanford spin off pioneering MIMO-OFDM (now Intel). Before joining EURECOM in 2004, he has been with the Department of Informatics, University of Oslo as an adjunct professor. D. Gesbert has published about 270 papers and 25 patents, some winning the 2015 IEEE Best Tutorial Paper Award (Communications Society), 2012 SPS Signal Processing Magazine Best Paper Award, 2004 IEEE Best Tutorial Paper Award (Communications Society), 2005 Young Author Best Paper Award for Signal Proc. Society journals, and paper awards at conferences 2011 IEEE SPAWC, 2004 ACM MSWiM. He is a Technical Co-chair for ICC2017 in Paris and WSA2017 in Berlin. He was named in the 2014 Thomson-Reuters List of Highly Cited Researchers in Computer Science. Since 2015, he holds the ERC Advanced grant PERFUME on the topic of smart device Communications in future wireless networks.