# Feedback Overhead Analysis for Base Station Cooperative Transmission

Xueying Hou, *Student Member, IEEE*, and Chenyang Yang, *Senior Member, IEEE*

***Abstract*—In this paper, we analyze the feedback overhead of channel direction information for downlink coherent base station (BS) cooperative transmission. The per-cell codebooks are considered, which are of practice importance. Instead of analyzing the required number of bits for feedback to keep a constant rate loss, we analyze the required overhead to ensure a target average signal-to-interference-plus-noise ratio (SINR) of each user. To this end, we formulate an optimization problem of bit allocation among the codebooks for local and cross channels that minimizes the total number of bits under the constraint of the average SINR, and find the explicit expression of the solution. We proceed to study the impact of various system parameters and channel features on the overall feedback overhead. Analytical and simulation results reveal that the overhead scales linearly with the overall number of transmit antennas, but decreases with the grow of the cell-edge signal-to-noise ratio. Moreover, the overhead can be significantly reduced by exploiting the diverse performance requirements of the users and the heterogeneous channel features through allocating the number of bits for feedback among multiple users and multiple per-cell links. This provides useful insight for the feedback strategy design of BS cooperation systems.**

***Index Terms*—Coordinated multi-point transmission, feedback overhead, per-cell codebook.**

## I. Introduction

BASE station (BS) cooperative transmission, called coordinated multi-point (CoMP) transmission in Long Term Evolution Advanced (LTE-A), is a promising strategy to provide high spectrum efficiency for cellular systems [1, 2]. Depending on the information shared among the BSs, CoMP systems can be divided into CoMP joint processing (CoMP-JP) (i.e., coherent BS cooperative transmission) and coordinated beamforming (CoMP-CB) [3, 4]. When both data and channel state information (CSI) are shared, CoMP-JP with multi-user (MU) multi-input multi-output (MIMO) precoding can exploit the full benefit of downlink CoMP systems.

To support MU-MIMO CoMP-JP transmission, the coordinated BSs need to obtain the downlink CSI for all users. In frequency division duplexing (FDD) systems, the required CSI at each BS is obtained through the uplink feedback of a codeword index with limited number of bits. The performance of single-cell limited feedback MU-MIMO systems has been extensively studied [5–7]. It was shown that the size of the codebook for quantizing the channel direction information (CDI) should grow linearly with the number of transmit antennas and signal-to-noise ratio (SNR) to ensure a constant rate loss. If we simply regard CoMP-JP as a single-cell MU-MIMO system with more antennas, both its multiplexing gain and feedback overhead will increase linearly with the number of cooperated BSs with respect to the corresponding Non-CoMP system. However, when the features of CoMP systems and channels are taken into account, it remains unclear how the feedback overhead scales with the system parameters, and whether the performance gain can justify the large feedback overhead.

CoMP-JP channel exhibits a unique structure, which is an aggregation of multiple single-cell channels from the cooperative BSs to each user. In general, the CoMP-JP channel is not independent and identically distributed (i.i.d.), and its statistics depends on the user's location. Moreover, the dimension of the CoMP-JP channel may vary in a system owing to the dynamic forming of the cooperative cluster. Although it is optimal to quantize the CoMP channel with a global codebook [8], a large number of optimal codebooks should be generated to accommodate the different channel statistics for the users in various locations. In practice, this is unrealistic since the complexity to generate the codebooks and to search the codewords are prohibited. Considering the unique features of the CoMP-JP channels and systems, it is natural to quantize the CoMP-JP channels with per-cell codebooks [9]. To provide a unified approach of codeword selection for CoMP-JP, CoMP-CB and Non-CoMP systems, the per-cell codewords are preferred to be selected independently. Although a feedback strategy with such structured codebooks and such a simple codeword selection is not optimal, it is scalable for large systems and of low complexity for implementation [9,10], and its performance can be improved by judiciously designing the codebook sizes for the per-cell codebooks. Furthermore, the well-designed single-cell codebooks can be reused, which is of practical importance for the system backward compatibility.

In this paper, we strive to quantify the overhead of CDI feedback for CoMP-JP precoding. All related works regarding the feedback overhead analysis for CoMP focused on ensuring a constant per-user rate loss, which actually concerns the spectral efficiency of the system. In [4,11], the scaling law of per-user feedback overhead of CoMP-CB system with either a single user or multiple users in each cell was analyzed, where a bit allocation among the desired and interference channels was considered. However, the overheads of CoMP-CB and CoMP-

Manuscript received April 17, 2012; revised October 9, 2012; accepted December 10, 2012. The associate editor coordinating the review of this paper and approving it for publication was H. Lin.

The authors are with the School of Electronics and Information Engineering, Beihang University (BUAA), Beijing, 100191, China (e-mail: hxymr@ee.buaa.edu.cn; cyyang@buaa.edu.cn).

This work was supported in part by the Key Project of Next Generation Wideband Wireless Communication Network of China under Grant No.2011ZX03003-001, and by the International S&T Cooperation Program of China (ISCP) under Grant No. 2008DFA12100.

Digital Object Identifier 10.1109/TWC.2013.013013.120534

JP systems differ, because CoMP-JP employs global CSI for joint precoding at the BSs and CoMP-CB employs per-cell CSI for individual precoding at each BS. In [9], the feedback overhead of CoMP-JP systems with per-cell codebooks was studied, where the codebook sizes for the per-cell CDIs are equal. It is worthy to note that except for providing high spectral efficiency, CoMP systems are expected to improve the performance for all users, especially when they are located at the cell edge. This implies that we can design the feedback strategy on demand, i.e., satisfying the requirement of each user, from another perspective.

To accommodate the system and channel features in practical CoMP systems, we will investigate how much feedback overhead is required to quantize the global CDI of each user to achieve its target performance. The contributions of this paper are in two folds:

1) We derive the scaling law of feedback overhead of CoMP-JP by finding the minimal number of bits for quantizing the CDI of each user under the constraint of the average single-to-interference-and-noise ratio (SINR), where the per-cell codebooks and independent codeword selection are considered. This is very different from the existing feedback overhead analysis as in [4, 5, 9, 11], where the scaling law was investigated to keep a constant rate loss.
2) We analyze the impact of various system parameters on the feedback overhead of CoMP-JP precoding, which sheds light upon the feedback strategy design for practical systems. Our analysis suggests that the feedback overhead of CoMP-JP will be dramatically reduced if the diverse performance requirements of multiple users and the non-i.i.d. feature of CoMP channels can be exploited.

The rest of the paper is organized as follows. Section II introduces the system model. Section III analyzes the feedback overhead. Simulations and conclusions are provided in Section IV and V.

## II. System Model

Consider a CoMP-JP system, where $N_b$ BSs each equipped with $N_t$ antennas cooperatively serve $M$ single-antenna users, and $M \leq N_b N_t$. We assume that the CSI from the coordinated BSs to the users are forwarded to a central unit (CU) via backhaul links with unlimited capacity and zero latency. When the backhaul is with limited capacity, other transmission strategies should be applied instead of CoMP-JP, e.g., CoMP-CB [3,4] or partial cooperation [12]. Therefore, different feedback schemes should be designed, which are not considered in this work. For simplicity, we refer CoMP-JP as CoMP in the rest of the paper.

The global channel from all the BSs to the $k$th user, $\text{MS}_k$, can be expressed as

$$\boldsymbol{g}_k = [\alpha_{k,1}\boldsymbol{h}_{k,1}^T, \ldots, \alpha_{k,N_b}\boldsymbol{h}_{k,N_b}^T]^T, \quad (1)$$

where $\alpha_{k,b} \in \mathbb{R}$ and $\boldsymbol{h}_{k,b} \in \mathbb{C}^{N_t \times 1}$ are respectively the large scale fading gain and small scale fading channel vector from the $b$th BS, i.e., $\text{BS}_b$, to $\text{MS}_k$, the entries of $\boldsymbol{h}_{k,b}$ are assumed as i.i.d. and zero-mean circularly symmetric complex Gaussian with unit variance, and $(\cdot)^T$ denotes the transpose. The BS who has the largest average channel gain with $\text{MS}_k$, say $\text{BS}_{b_k}$, is its local BS and all the other $N_b - 1$ BSs are its cooperative BSs. The channel from $\text{BS}_{b_k}$ to $\text{MS}_k$ is called local channel, and the channels from the cooperative BSs to $\text{MS}_k$ are referred to as cross channels.

The data to be transmitted to all the users are $\boldsymbol{d} = [d_1, \ldots, d_M]^T$. Without loss of generality, we assume that $\mathbb{E}\{\boldsymbol{d}\boldsymbol{d}^H\} = \mathbf{I}_M$, where $\mathbb{E}\{\cdot\}$ is the expectation operator, $(\cdot)^H$ denotes the conjugate transpose, and $\mathbf{I}_M$ represents an identity matrix of dimension $M$. The received signal at $\text{MS}_k$ is

$$y_k = \boldsymbol{g}_k^H \boldsymbol{v}_k \sqrt{p_k} d_k + \sum\nolimits_{j=1, j\neq k}^{M} \boldsymbol{g}_k^H \boldsymbol{v}_j \sqrt{p_j} d_j + n_k, \quad (2)$$

where $\boldsymbol{v}_j \in \mathbb{C}^{N_b N_t \times 1}$ and $p_j$ respectively represent the unit-norm beamforming vector and the power allocated from all the cooperative BSs to $\text{MS}_j$, $j = 1, \ldots, M$, and $n_k$ is additive white Gaussian noise (AWGN) with zero mean and variance $\sigma_k^2$.

### A. Finite Rate Feedback Model

Assume that each user has perfect knowledge of its global channel vector. We consider the *per-cell codebook* based limited feedback [9, 10], where each user (say, $\text{MS}_k$) employs single-cell codebooks to separately quantize its per-cell channels, $\boldsymbol{h}_{k,b}$, $b = 1, \ldots, N_b$. We assume that the instantaneous norms of per-cell channels $\|\boldsymbol{h}_{k,b}\|$, $k = 1, \ldots, M$, $b = 1, \ldots, N_b$, are perfectly fed back, where $\|\cdot\|$ represents the two-norm. Since the per-cell large scale channel gains $\alpha_{k,b}$, $b = 1, \ldots, N_b$, are slow varying variables, we assume that they can be obtained at $\text{MS}_k$ by averaging the received signals over a certain period[1] and then fed back to the BSs. After $\text{MS}_k$ quantizes each per-cell CDI $\bar{\boldsymbol{h}}_{k,b} = \boldsymbol{h}_{k,b}/\|\boldsymbol{h}_{k,b}\|$, it feeds back the quantized version to its local BS. The cooperative BSs forward their gathered CSI to the CU, who finally reconstructs the global channels for all users.

When the per-cell codebook is employed, the per-cell codewords can be selected either jointly to minimize the quantization error of the global CDI, or independently to minimize the quantization error of each per-cell CDI. We consider independent codeword selection [10], where the quantized per-cell CDI is obtained at $\text{MS}_k$ as

$$\hat{\boldsymbol{h}}_{k,b} = \arg \max_{\boldsymbol{c}_j \in \mathcal{C}_{k,b}} |\bar{\boldsymbol{h}}_{k,b}^H \boldsymbol{c}_j|, \quad (3)$$

which performs fairly well when phase ambiguity (PA) information, $e^{j\phi_{k,b}} \triangleq \hat{\boldsymbol{h}}_{k,b}^H \bar{\boldsymbol{h}}_{k,b} / |\hat{\boldsymbol{h}}_{k,b}^H \bar{\boldsymbol{h}}_{k,b}|$, is fed back. $\mathcal{C}_{k,b}$ is the per-cell codebook, which consists of unit norm codewords $\boldsymbol{c}_j \in \mathbb{C}^{N_t \times 1}$, $j = 1, \ldots, 2^{B_{k,b}^{\text{CDI}}}$, and $B_{k,b}^{\text{CDI}}$ is the number of bits used to quantize the per-cell CDI $\bar{\boldsymbol{h}}_{k,b}$.

The PA is a phase rotation between each per-cell CDI and each per-cell codeword. It does not affect the performance of single-cell limited feedback MIMO systems [13], but affects the performance of CoMP-JP systems with per-cell codebooks, especially when independent codeword selection is applied [10]. This is because the codeword selection in (3) only maximizes the magnitude of $\bar{\boldsymbol{h}}_{k,b}^H \boldsymbol{c}_j$, but ignores the phase of it. As a result, even when the quantization

[1]The period depends on the variation of the location and shadowing of a mobile user.

of per-cell CDIs are perfect, the quantized per-cell CDIs will be phase-rotated versions of the per-cell CDIs, i.e., $\hat{\boldsymbol{h}}_{k,b} = \bar{\boldsymbol{h}}_{k,b} e^{-j\phi_{k,b}}$, $b = 1, \ldots, N_b$. If the random phases $\phi_{k,b}$, $b = 1, \ldots, N_b$, are not fed back, after receiving the quantized per-cell CDIs the CU reconstructs the global channel vector as $[\alpha_{k,1}\|\boldsymbol{h}_{k,1}\|\bar{\boldsymbol{h}}_{k,1}^T e^{-j\phi_{k,1}}, \ldots, \alpha_{k,N_b}\|\boldsymbol{h}_{k,N_b}\|\bar{\boldsymbol{h}}_{k,N_b}^T e^{-j\phi_{k,N_b}}]^T$, which will differ from the true value of the global channel vector shown in (1) and will lead to multiplicative quantization error of the global CDI that hinders the co-phasing of the coherent CoMP transmission. If the PA differences $\omega_{k,b} \triangleq \phi_{k,b} - \phi_{k,b_k}, b = 1, \ldots, N_b, b \neq b_k$, can be fed back with a few bits after scalar quantization, however, the CoMP system will even outperform that using the joint codeword selection without PA feedback [10].

After $\text{MS}_k$ quantizes the per-cell CDIs and the PA differences, it feeds back the indices of the selected codewords to its local BS. Then all cooperative BSs send the gathered CSI to the CU, who reconstructs the quantized version of the global channel of $\text{MS}_k$ as follows

$$\hat{\boldsymbol{g}}_k = \Big[\alpha_{k,1}\|\boldsymbol{h}_{k,1}\|\hat{\boldsymbol{h}}_{k,1}^T e^{j\hat{\omega}_{k,1}}, \ldots, \alpha_{k,b_k}\|\boldsymbol{h}_{k,b_k}\|\hat{\boldsymbol{h}}_{k,b_k}^T, \ldots, \alpha_{k,N_b}\|\boldsymbol{h}_{k,N_b}\|\hat{\boldsymbol{h}}_{k,N_b}^T e^{j\hat{\omega}_{k,N_b}}\Big]^T, \quad (4)$$

where $\hat{\omega}_{k,b}$ is the quantized PA difference, $b = 1, \ldots, N_b, b \neq b_k$, with which the PA can be compensated.

Define $\sin^2\theta_{k,b} = 1 - |\bar{\boldsymbol{h}}_{k,b}^H \hat{\boldsymbol{h}}_{k,b}|^2$, which is the instantaneous per-cell CDI quantization error. We consider random vector quantization (RVQ) for the tractability of analysis. Any well-designed codebooks will be superior to the RVQ codebook. From [5], the per-cell CDI can be expressed as

$$\bar{\boldsymbol{h}}_{k,b} = \cos\theta_{k,b} e^{j\phi_{k,b}} \hat{\boldsymbol{h}}_{k,b} + \sin\theta_{k,b} \boldsymbol{s}_{k,b}, \quad (5)$$

where $\boldsymbol{s}_{k,b} \in \mathbb{C}^{N_t \times 1}$ is a unit norm vector isotropically distributed in the null space of $\hat{\boldsymbol{h}}_{k,b}$ [5], and $\phi_{k,b}$ is the PA uniformly distributed within $[-\pi, \pi]$ [13].

Then from (1) and (4), the global channel can be expressed as

$$\boldsymbol{g}_k = e^{j\phi_{k,b_k}} \boldsymbol{D}_k \hat{\boldsymbol{g}}_k + \boldsymbol{s}_{\boldsymbol{g}_k}, \quad (6)$$

where $\boldsymbol{D}_k = \text{diag}\{\cos\theta_{k,1} e^{j\Delta\omega_{k,1}} \mathbf{I}_{N_t}, \ldots, \cos\theta_{k,b_k} \mathbf{I}_{N_t}, \ldots, \cos\theta_{k,N_b} e^{j\Delta\omega_{k,N_b}} \mathbf{I}_{N_t}\}$ is a multiplicative error that comes from the quantization errors of the PA differences $\Delta\omega_{k,b} = \omega_{k,b} - \hat{\omega}_{k,b}$, $b = 1, \ldots, N_b$, $b \neq b_k$, $\text{diag}\{\cdot\}$ is diagonalization operation, and $\boldsymbol{s}_{\boldsymbol{g}_k} = [\alpha_{k,1}\sin\theta_{k,1}\|\boldsymbol{h}_{k,1}\|\boldsymbol{s}_{k,1}^T, \ldots, \alpha_{k,N_b}\sin\theta_{k,N_b}\|\boldsymbol{h}_{k,N_b}\|\boldsymbol{s}_{k,N_b}^T]^T$ is an additive error that is caused by the quantization error of the per-cell CDIs.

Since RVQ is used to quantize the per-cell CDI, we have $\mathbb{E}\{\sin^2\theta_{k,b}\} \leq 2^{\frac{-B_{k,b}^{\text{CDI}}}{N_t - 1}}$ [5]. When uniform quantization is applied to quantize the PA difference, we have $\mathbb{E}\{\Delta\omega_{k,b}\} = 0$ and $\mathbb{E}\{|\Delta\omega_{k,b}|^2\} = \frac{\pi^2}{3} 2^{-2B_{k,b}^{\text{PA}}}$ [8], where $B_{k,b}^{\text{PA}}$ is the number of bits used by $\text{MS}_k$ to quantize $\omega_{k,b}$. Considering that $\alpha_{k,b}\|\boldsymbol{h}_{k,b}\|$, $b = 1, \ldots, N_b, k = 1, \ldots, M$, are known at the cooperative BSs, the CU can compute the statistics of $\boldsymbol{s}_{\boldsymbol{g}_k}$ and $\boldsymbol{D}_k$ with $B_{k,b}^{\text{CDI}}$ and $B_{k,b}^{\text{PA}}$, and then the CU can estimate the SINR of $\text{MS}_k$.

The feature of CoMP channels implies that different users need different numbers of bits for quantization to achieve the same channel quality. It is worthy to note that the distortion of the received CDI at the BSs also depends on the condition of feedback channels, which differ for multiple users [7]. This suggests that the uplink bandwidth for each user to feed back the same number of bits will differ if the user is allowed to select preferred modulation and coding according to its uplink SNR. Nonetheless, for easy implementation, prevalent systems simply apply low-rate coding and low-order modulation to ensure the reliability of the channel feedback, no matter if the feedback channel is in good condition. As a first attempt to analyze the minimal feedback overhead for CoMP, we restrict ourselves to analyze the impact of the channel quantization, i.e., the required number of bits, rather than the required uplink bandwidth for feedback. The assumption behind such an analysis is error-free uplink transmission for feedback.

### *B. Multi-cell Scheduling and Precoding*

With the reconstructed global channels of all users, the CU selects $M$ users to serve in the same time-frequency resource with multi-cell MU-MIMO precoding. Spatial scheduling has a large impact on the performance of MU-MIMO systems by increasing the received signal power as well as by improving the performance of the precoding. In limited feedback systems, such a gain of user selection largely depends on the channel quality information (CQI) and incurs huge overhead [14]. For CoMP systems, it is an on-going research topic about how to reduce the feedback overhead for scheduling. One possibility is using a two-stage feedback [15], where all candidate users feed back a rough channel information for scheduling in the first stage and only the selected users feed back a refined channel information for beamforming in the second stage. Another natural way is only serving the cell-edge users with CoMP thereby only these users need to feed back their channels to multiple BSs for multi-cell scheduling and precoding. When these practical feedback schemes are applied, one benefit of the scheduling, i.e., improving the received signal power by selecting the users with large channel gains, will lose in a large extent, while the other benefit of scheduling, i.e., improving the precoding performance by choosing the users with semi-orthogonal CDIs, will largely remain. Therefore, as many related works in the literature [4, 5, 7, 9], we decouple the impact of multiuser scheduling and precoding, and only study the feedback overhead of the global CDI for precoding. To this end, we do not consider any specific scheduling criterion. To show the impact of scheduling on the performance of precoding, we will introduce a parameter to reflect the orthogonality among the users as shown later.

We consider a low complexity precoding structure, which consists of a pseudo-inverse based zero forcing beamforming (ZFBF) and a power allocation. Under sum power constraint, such a ZFBF in conjunction with an optimal power allocation that maximizes the sum rate has been proved to achieve the maximal sum rate [16]. Under per-BS power constraint (PBPC) in CoMP, this zero forcing (ZF) precoding has minor performance loss from the optimal ZF precoding when there are many users in each cell [17]. However, its performance is very hard to analyze. In this paper, we consider a sub-optimal but more tractable power constraint, which is the

per-user power constraint (PUPC) [18]. It has been shown that the ZFBF with power allocation under different power constraints perform closely for large number of users [17]. We assume that the transmit powers of the BSs are all equal to $P_0$. Then the power allocated to MS$_k$ is $p_k = N_b P_0/M$. Denote $\hat{\boldsymbol{G}} = [\hat{\boldsymbol{g}}_1, \ldots, \hat{\boldsymbol{g}}_M]^T$ and $\boldsymbol{V} = \hat{\boldsymbol{G}}^H(\hat{\boldsymbol{G}}\hat{\boldsymbol{G}}^H)^{-1}$, then the beamforming vector for MS$_k$ is $\boldsymbol{v}_k = \boldsymbol{V}(:,k)/\|\boldsymbol{V}(:,k)\|$, where $\boldsymbol{V}(:,k)$ denotes the $k$th column of matrix $\boldsymbol{V}$. We will verify in Section IV by simulation that the feedback overhead with PUPC is close to that with PBPC.

## III. Feedback Overhead Analysis of CoMP Transmission

In this section, we first find the minimal overhead required by each user to support a target quality of service (QoS). Then we analyze the overall feedback overhead of the FDD CoMP systems.

### A. Minimal Feedback Overhead for Each User to Achieve the Required QoS

A well-known approach of analyzing the feedback overhead is to find a connection of the average per-user rate loss with the codebook size [5]. However, no closed-form expression is available for the ergodic per-user rate even for single-cell MU-MIMO systems. For i.i.d. channels, an upper bound of the per-user rate loss was provided in [5]. Unfortunately, the derived upper bound can not be extended to CoMP systems whose channels are no longer i.i.d.. On the other hand, the per-user rate loss is not able to be achieved in a real-world system, because it was derived under an implicit assumption: the BS and each user have the same knowledge of the SINR. In MU-MIMO FDD systems, each user knows its own CSI but is unaware of the precoding vectors for other users, while the BS knows all the precoding vectors but does not see the same CSI as the users. As a result, neither each user nor the BS can accurately compute the downlink SINR experienced by the user, and the user and BS "see" different SINRs. This is a fundamental challenge for designing limited feedback MU-MIMO systems. In practice, the BS needs to estimate the SINR with the quantized CSI as the CQI for scheduling multiple users [14] and for adjusting their downlink transmission rates.

In this paper we employ an alternative approach to analyze the overhead. Instead of finding the relationship of the per-user rate loss and the codebook size, we derive the required minimal feedback overhead of each user to achieve a target SINR. To exploit the non-i.i.d. feature of CoMP channels and the diverse QoS requirements of multiple users, we allow the users to employ different overall number of bits for quantizing their global CDIs and different codebook sizes for their local and cross CDIs. Since the codebook sizes should not change too frequently in case causing a large signaling overhead, we find the minimal numbers of bits to achieve an average SINR.

We begin with estimating the SINR of each user at the CU. The received power at MS$_k$ of the signals transmitted from the cooperative BSs to MS$_j$ can be obtained from (2) as $q_{k,j} = \frac{N_b P_0}{M}|\boldsymbol{g}_k^H \boldsymbol{v}_j|^2$. When $j = k$, $q_{k,k}$ is the received power of the desired signal, otherwise it is the received interference power. As we have explained, the CU can obtain the statistics of the channel quantization errors $\boldsymbol{D}_k$ and $\boldsymbol{s}_{\boldsymbol{g}_k}$ with the knowledge of numbers of bits for quantizing the per-cell CDIs and the PA differences. Considering that both the quantized global channel vectors $\hat{\boldsymbol{g}}_k$ and the precoding vectors $\boldsymbol{v}_j$ are available at the CU, the CU can obtain the minimum mean square error (MMSE) estimate of the received power as

$$\begin{aligned}\hat{q}_{k,j}^{\mathrm{MMSE}} =& \frac{N_b P_0}{M}\mathbb{E}_{\boldsymbol{D}_k,\boldsymbol{s}_{\boldsymbol{g}_k},\phi_{k,b_k}}\left\{\left|e^{-j\phi_{k,b_k}}\hat{\boldsymbol{g}}_k^H\boldsymbol{D}_k^H\boldsymbol{v}_j + \boldsymbol{s}_{\boldsymbol{g}_k}^H\boldsymbol{v}_j\right|^2\middle|\boldsymbol{v}_j,\hat{\boldsymbol{g}}_k\right\}\\ =& \frac{N_b P_0}{M}\mathbb{E}_{\boldsymbol{D}_k}\left\{\left|\hat{\boldsymbol{g}}_k^H\boldsymbol{D}_k^H\boldsymbol{v}_j\right|^2\right\} + \frac{N_b P_0}{M}\mathbb{E}_{\boldsymbol{s}_{\boldsymbol{g}_k}}\left\{\left|\boldsymbol{s}_{\boldsymbol{g}_k}^H\boldsymbol{v}_j\right|^2\right\},\end{aligned} \tag{7}$$

where $\mathbb{E}_x\{\cdot\}$ represents the expectation over a random variable $x$, the first equality comes by substituting (6) into the expression of $q_{k,j}$, and the second equality is derived by averaging over $\phi_{k,b_k}$.

Then the instantaneous received signal power and interference power of MS$_k$ can be estimated at the CU as $\hat{\mathrm{S}}_k = \hat{q}_{k,k}^{\mathrm{MMSE}}$ and $\hat{\mathrm{I}}_k = \sum_{j=1,j\neq k}^{M}\hat{q}_{k,j}^{\mathrm{MMSE}}$, and the instantaneous receive SINR of MS$_k$ can be estimated as $\hat{\gamma}_k = \frac{\hat{\mathrm{S}}_k}{\hat{\mathrm{I}}_k+\sigma_k^2}$.[2] Denote the target SINR of MS$_k$ as $\gamma_k^{\mathrm{QoS}}$. To remove the impact of small scale fading channels on the feedback overhead, we consider an average SINR requirement of MS$_k$, i.e., $\mathbb{E}\{\hat{\gamma}_k\} \geq \gamma_k^{\mathrm{QoS}}$.

*Proposition* 1. The optimization problem that minimizes the total number of bits for quantizing the global CDI of MS$_k$ under the average SINR constraint can be formulated as

$$\min_{B_{k,b}^{\mathrm{CDI}},B_{k,b}^{\mathrm{PA}}} \sum\nolimits_{b=1}^{N_b} B_{k,b}^{\mathrm{CDI}} + \sum\nolimits_{b=1,b\neq b_k}^{N_b} B_{k,b}^{\mathrm{PA}} \tag{8a}$$

$$\text{s.t.} \quad C_k - \sum\nolimits_{b=1}^{N_b} A_{k,b}^{\mathrm{CDI}} 2^{\frac{-B_{k,b}^{\mathrm{CDI}}}{N_t-1}} - \sum\nolimits_{b=1,b\neq b_k}^{N_b} A_{k,b}^{\mathrm{PA}} 2^{-2B_{k,b}^{\mathrm{PA}}} \geq 0 \tag{8b}$$

$$B_{k,b}^{\mathrm{CDI}} \geq 0,\ b = 1,\ldots,N_b, \tag{8c}$$

$$B_{k,b}^{\mathrm{PA}} \geq 0,\quad b = 1,\ldots,N_b, b \neq b_k, \tag{8d}$$

and the minimal numbers of bits for quantizing each per-cell CDI and each PA difference are

$$B_{k,b}^{\mathrm{CDI}} = (N_t-1)\left[\log_2 A_{k,b}^{\mathrm{CDI}} - \log_2(N_t-1)\lambda_k\right]^+, \tag{9a}$$

$$B_{k,b}^{\mathrm{PA}} = \frac{1}{2}\left[\log_2 A_{k,b}^{\mathrm{PA}} - \log_2\frac{1}{2}\lambda_k\right]^+,\ b \neq b_k, \tag{9b}$$

where $A_{k,b}^{\mathrm{CDI}} = (\mu_k N_t + \gamma_k^{\mathrm{QoS}}\sum_{j=1,j\neq k}^{M}\bar{\beta}_{j,b})\alpha_{k,b}^2$, $A_{k,b}^{\mathrm{PA}} = N_t\frac{\pi^2}{3}(\mu_k + \gamma_k^{\mathrm{QoS}})(1-\bar{\beta}_{k,b})\alpha_{k,b}^2$, $\bar{\beta}_{k,b} = \frac{\alpha_{k,b}^2}{\sum_{l=1}^{N_b}\alpha_{k,l}^2}$, $C_k = \mu_k N_t\sum_{l=1}^{N_b}\alpha_{k,l}^2 - \gamma_k^{\mathrm{QoS}}\frac{M\sigma_k^2}{N_b P_0}$ is a constant independent from the codebook size, $\mu_k = \mathbb{E}\{|\boldsymbol{v}_k^H\frac{\hat{\boldsymbol{g}}_k}{\|\hat{\boldsymbol{g}}_k\|}|^2\}$ reflects the average

[2] $\sigma_k^2$ includes the power of noise and that of the other-cluster interference, which is assumed available at the CU. It can be obtained by long term interference measurement at MS$_k$ and then be fed back to the BSs with negligible feedback overhead.

orthogonality between the channel of $\mathrm{MS}_k$ and the subspace spanned by the channels of its co-scheduled users, a large value of $\mu_k$ indicates a better orthogonality, $\lambda_k$ is the Lagrangian multiplier that should satisfy (8b), and $[x]^+ = \max\{0, x\}$.

*Proof:* See Appendix A. ■

(9a) and (9b) are in fact a bit allocation among the per-cell codebooks and among the PA differences, which is similar to the water-filling power allocation. $A_{k,b}^{\mathrm{CDI}}$ reflects the impact of the quantization error of the per-cell CDI from $\mathrm{BS}_b$ to $\mathrm{MS}_k$ $\bar{\boldsymbol{h}}_{k,b}$. When its value is large, more bits should be allocated to quantize $\bar{\boldsymbol{h}}_{k,b}$. The same is true for the PA difference quantization. Moreover, we can observe that the number of bits depends on the value of $\alpha_{k,b}$ as well as the values of $\bar{\beta}_{k,b}$ and $\bar{\beta}_{j,b}$ $(j \neq k)$, which respectively reflect the imbalance of average per-cell channel gains of $\mathrm{MS}_k$ and its co-scheduled users.

It is worthy to note that all the parameters required for the bit allocation only depend on the average channel information (i.e., the locations of the desired and co-scheduled users) and the system configuration (i.e., the number of cooperative BSs, the number of antennas at each BS, the SNR, and the target average SINR of each user). This suggests that in practice the bit allocation can be implemented semi-dynamically. A detailed procedure of the bit allocation can be found in Section IV of [19], though the optimization problem there is different from us.

To analyze how the feedback overhead scales with the system parameters, in the following we find an explicit expression of the total number of feedback bits by removing the Lagrangian multiplier. Consider $B_{k,b}^{\mathrm{CDI}} > 0$ and $B_{k,b}^{\mathrm{PA}} > 0$. This corresponds to the scenario where the cooperative set of BSs for $\mathrm{MS}_k$ is well designed, such that all the per-cell CDIs and PA differences need to be fed back. Then by substituting (9a) and (9b) into (8b) and after some regular manipulations, we can obtain $\lambda_k = \frac{C_k}{N_b(N_t-1/2)-1/2}$.

Note that the target SINR should not exceed the maximal average SINR with perfect CSI, otherwise it will not be achievable. When perfect CSI is available at the BSs, the MUI can be avoided by ZFBF, and the maximal SINR can be achieved when the scheduled users are mutually orthogonal. Define the cell-edge SNR as $\mathrm{SNR}_{\mathrm{edge}} \triangleq \frac{P_0}{\sigma_0^2}\alpha_{\mathrm{edge}}^2$, where $\alpha_{\mathrm{edge}}$ and $\sigma_0^2$ are the large scale fading gain and the noise variance of the cell-edge user, respectively. The cell-edge SNR reflects the density of a cellular network. A Non-CoMP system with higher cell-edge SNR is more interference-limited, hence the corresponding CoMP system will provide higher performance gain. Then the maximum average SINR of $\mathrm{MS}_k$ with perfect CSI can be obtained as

$$\begin{aligned}\gamma_k^{\mathrm{Perf}} &\triangleq \mathbb{E}\left\{\frac{N_bP_0}{M}\frac{\|\boldsymbol{g}_k\|^2}{\sigma_k^2}\right\} = N_t\frac{N_bP_0}{M\sigma_k^2}\sum\nolimits_{l=1}^{N_b}\alpha_{k,l}^2 \\ &= \left(N_t\frac{N_b}{M}\frac{\sigma_0^2}{\sigma_k^2}\sum_{l=1}^{N_b}\frac{\alpha_{k,l}^2}{\alpha_{\mathrm{edge}}^2}\right)\mathrm{SNR}_{\mathrm{edge}}. \end{aligned} \tag{10}$$

Substituting the expression of $\lambda_k$ into (9a) and (9b) and considering (10), we can obtain the total numbers of bits respectively for quantizing the per-cell CDIs and PA differences of $\mathrm{MS}_k$ as follows

$$\begin{aligned} B_{k,\mathrm{sum}}^{\mathrm{CDI}} &= \sum\nolimits_{b=1}^{N_b} B_{k,b}^{\mathrm{CDI}} \\ &= \underbrace{(N_t-1)\sum_{b=1}^{N_b}\log_2\left(1+\frac{1}{N_t}\frac{1}{\mu_k}\gamma_k^{\mathrm{QoS}}\sum_{j=1,j\neq k}^{M}\bar{\beta}_{j,b}\right)}_{B_{k,\mathrm{sum}}^{\mathrm{CDI}}(1)} \\ &\quad + \underbrace{N_b(N_t-1)\log_2\frac{\left(\prod_{l=1}^{N_b}\alpha_{k,l}^2\right)^{1/N_b}}{\frac{1}{N_b}\sum_{l=1}^{N_b}\alpha_{k,l}^2}}_{B_{k,\mathrm{sum}}^{\mathrm{CDI}}(2)} + C_0^{\mathrm{CDI}} \\ &\quad + \underbrace{N_b(N_t-1)\log_2\left(\frac{1}{1-\frac{1}{\mu_k}\gamma_k^{\mathrm{QoS}}/\gamma_k^{\mathrm{Perf}}}\right)}_{B_{k,\mathrm{sum}}^{\mathrm{CDI}}(3)}, \end{aligned} \tag{11}$$

$$\begin{aligned} B_{k,\mathrm{sum}}^{\mathrm{PA}} &= \sum\nolimits_{b=1,b\neq b_k}^{N_b} B_{k,b}^{\mathrm{PA}} \\ &= \underbrace{(N_b-1)\frac{1}{2}\log_2\left[\left(1+\frac{1}{\mu_k}\gamma_k^{\mathrm{QoS}}\right)\right]}_{B_{k,\mathrm{sum}}^{\mathrm{PA}}(1)} \\ &\quad + \underbrace{\frac{1}{2}\sum_{b=1,b\neq b_k}^{N_b}\log_2(1-\bar{\beta}_{k,b})\bar{\beta}_{k,b}}_{B_{k,\mathrm{sum}}^{\mathrm{PA}}(2)} + C_0^{\mathrm{PA}} \\ &\quad + \underbrace{(N_b-1)\frac{1}{2}\log_2\left(\frac{1}{1-\frac{1}{\mu_k}\gamma_k^{\mathrm{QoS}}/\gamma_k^{\mathrm{Perf}}}\right)}_{B_{k,\mathrm{sum}}^{\mathrm{PA}}(3)}, \end{aligned} \tag{12}$$

where $C_0^{\mathrm{CDI}} = N_b(N_t-1)\log_2\frac{N_b(N_t-\frac{1}{2})-\frac{1}{2}}{N_b(N_t-1)}$ and $C_0^{\mathrm{PA}} = (N_b-1)\frac{1}{2}\log_2\{\frac{\pi^2}{3}[N_b(2N_t-1)-1]\}$ are constants for a given system.

The minimal feedback overhead of $\mathrm{MS}_k$ to achieve the target SINR is $B_{k,\mathrm{sum}} = B_{k,\mathrm{sum}}^{\mathrm{CDI}} + B_{k,\mathrm{sum}}^{\mathrm{PA}}$.

### *B. Feedback Overhead Analysis for CoMP Transmission*

In the following, we study the impact of CoMP channels and system parameters on the feedback overhead. We will emphasize the difference from single-cell systems, which comes from the fact that the CoMP channels are non-i.i.d. and the per-cell codebooks are employed.

*1) Impact of the Numbers of Co-scheduled Users, Cooperative BSs and Transmit Antennas :* To connect with existing results in single-cell MU-MIMO systems, we consider a group of special users whose large scale fading gains are identical, i.e., $\alpha_{k,1} = \cdots = \alpha_{k,N_b} \triangleq \alpha_{\mathrm{edge}}$, $k = 1, \ldots, M$. Since we have assumed i.i.d. per-cell channels, we know from (1) that the global channels of these users are also i.i.d.. We refer to these users as the "exact cell-edge users".[3]

For an "exact cell-edge user" $\mathrm{MS}_k$, the value of the term $B_{k,\mathrm{sum}}^{\mathrm{CDI}}(2)$ in (11) reduces to zero. Moreover, we have

[3]Note that this is a mathematical definition to facilitate the comparison with single-cell systems. In practice, when considering the path loss, shadowing and sector antenna power gains, such a user may not be located at the exact cell edge and may rarely appear.

$\bar{\beta}_{k,b} = \frac{\alpha_{k,b}^2}{\sum_{l=1}^{N_b} \alpha_{k,l}^2} = \frac{1}{N_b}$, $b = 1, \ldots, N_b$, $k = 1, \ldots, M$, under which $B_{k,\mathrm{sum}}^{\mathrm{CDI}}(1) = N_b(N_t - 1)\log_2(1 + \gamma_k^{\mathrm{QoS}} \frac{M-1}{N_t N_b \mu_k})$ and $B_{k,\mathrm{sum}}^{\mathrm{PA}}(2) = (N_b - 1)\frac{1}{2}\log_2 \frac{N_b - 1}{N_b^2}$. By substituting these terms into (11) and (12), the total number of feedback bits of $\mathrm{MS}_k$ can be obtained as

$$\begin{aligned} B_{k,\mathrm{sum}} =& B_{k,\mathrm{sum}}^{\mathrm{CDI}} + B_{k,\mathrm{sum}}^{\mathrm{PA}} \\ =& N_b(N_t - 1)\log_2\left(1 + \gamma_k^{\mathrm{QoS}} \frac{M-1}{N_t N_b \mu_k}\right) \\ &+ (N_b - 1)\frac{1}{2}\log_2\left(1 + \frac{1}{\mu_k}\gamma_k^{\mathrm{QoS}}\right) \\ &+ \left(N_b N_t - \frac{1}{2}N_b - \frac{1}{2}\right)\log_2\left(\frac{1}{1 - \frac{1}{\mu_k}\gamma_k^{\mathrm{QoS}}/\gamma_k^{\mathrm{Perf}}}\right) \\ &+ C_{\mathrm{edge}}, \end{aligned} \tag{13}$$

where $C_{\mathrm{edge}} = C_0^{\mathrm{CDI}} + B_{k,\mathrm{sum}}^{\mathrm{PA}}(2) + C_0^{\mathrm{PA}} = N_b(N_t - 1)\log_2 \frac{N_b(N_t - \frac{1}{2}) - \frac{1}{2}}{N_b(N_t - 1)} + (N_b - 1)\frac{1}{2}\log_2\left[\frac{N_b - 1}{N_b}\frac{\pi^2}{3}(2N_t - 1 - \frac{1}{N_b})\right]$ is a constant for a given system.

From (13) we can observe that, since we do not assume full spatial multiplexing as in [5], the overhead we obtained increases with the number of co-scheduled users $M$ for a given target SINR. This is because when $M$ increases, the power allocated to $\mathrm{MS}_k$ will be reduced due to the power constraint at the BSs and the PUPC. At the same time, the interference generated to $\mathrm{MS}_k$ will increase. To achieve the target SINR, more bits should be allocated to the users for reducing the MUI.

To analyze the scaling law of the feedback overhead from (13) and connect with the single-cell result derived in [5], define $R_k^{\mathrm{Loss}} \triangleq \log_2(1 + \mu_k \gamma_k^{\mathrm{Perf}}) - \log_2(1 + \gamma_k^{\mathrm{QoS}}) \approx \log_2\left(\frac{\mu_k \gamma_k^{\mathrm{Perf}}}{\gamma_k^{\mathrm{QoS}}}\right)$, where the approximation is accurate when the values of $\gamma_k^{\mathrm{QoS}}$ and $\gamma_k^{\mathrm{Perf}}$ are large.[4] Then, $\gamma_k^{\mathrm{QoS}} = \mu_k \gamma_k^{\mathrm{Perf}} / 2^{R_k^{\mathrm{Loss}}}$. Note that $R_k^{\mathrm{Loss}}$ is not a rate loss. Nonetheless, it is an upper bound of the average rate loss considering the Jensen's inequity. We further consider full multiplexing as in [5], i.e., $M = N_b N_t$. Then for the high target SINR $\gamma_k^{\mathrm{QoS}}$, the feedback overhead of $\mathrm{MS}_k$ can be approximated as

$$\begin{aligned} B_{k,\mathrm{sum}} \approx& \left(N_b N_t - \frac{1}{2}N_b - \frac{1}{2}\right)\log_2\left(\frac{\frac{1}{\mu_k}\gamma_k^{\mathrm{QoS}}}{1 - \frac{1}{\mu_k}\gamma_k^{\mathrm{QoS}}/\gamma_k^{\mathrm{Perf}}}\right) \\ &+ N_b(N_t - 1)\log_2\left(\frac{N_b N_t - 1}{N_b N_t}\right) + C_{\mathrm{edge}} \end{aligned} \tag{14}$$

$$\begin{aligned} =& \left(N_b N_t - \frac{1}{2}N_b - \frac{1}{2}\right)\log_2\left(\frac{\gamma_k^{\mathrm{Perf}}}{2^{R_k^{\mathrm{Loss}}} - 1}\right) \\ &+ N_b(N_t - 1)\log_2\left(\frac{N_b N_t - 1}{N_b N_t}\right) + C_{\mathrm{edge}} \end{aligned} \tag{15}$$

$$\begin{aligned} =& \left(N_b N_t - \frac{1}{2}N_b - \frac{1}{2}\right)\left[\frac{\mathrm{SNR}_{\mathrm{edge}}^{\mathrm{dB}} + 10\log_{10} N_b}{3}\right. \\ &\left. - \log_2(2^{R_k^{\mathrm{Loss}}} - 1)\right] + C_{\mathrm{CoMP}}, \end{aligned} \tag{16}$$

[4]High $\gamma_k^{\mathrm{Perf}}$ was implicitly assumed for the scaling law analysis in [5] because high power region was considered. To achieve a constant rate loss, $\gamma_k^{\mathrm{QoS}}$ should also be high.

where $\mathrm{SNR}_{\mathrm{edge}}^{\mathrm{dB}} = 10\log_{10}\mathrm{SNR}_{\mathrm{edge}}$, (16) is obtained from substituting (10) into (15), and $C_{\mathrm{CoMP}} = N_b(N_t - 1)\log_2\left(\frac{N_b N_t - 1}{N_b N_t}\right) + C_{\mathrm{edge}}$ is a constant for a given system.

It follows that the per-user feedback overhead of CoMP systems needs to scale linearly with the cell-edge SNR to keep a constant value of $R_k^{\mathrm{Loss}}$, and the scaling slope is $N_b N_t - \frac{1}{2}N_b - \frac{1}{2}$. According to the analysis for single-cell MU-MIMO systems [5, 20], for a channel vector with $N_b N_t$ complex dimensions, the scaling slope is $N_b N_t - 1$ because the norm and the PA of the channel vector are unnecessary for ZFBF. The reduction of the slope $\frac{N_b - 1}{2}$ in CoMP comes from the fact that we do not consider the overhead to quantize the $N_b$ per-cell channel norms, each of which is of $\frac{1}{2}$ complex dimension, and one PA in CoMP is unnecessary to quantize such that $\frac{1}{2}$ complex dimension is subtracted. The term $10\log_{10} N_b$ in the first term within the bracket reflects the increased cell-edge SNR led by the coherent cooperative transmission of the $N_b$ BSs. The term $C_{\mathrm{CoMP}}$ is an extra overhead compared with the single-cell results, which is induced by the per-cell codebook. Under asymptotic large regime of $N_t$, $C_{\mathrm{CoMP}} \approx (N_b - 1)\frac{1}{2}\log_2(\frac{2\pi^2}{3}N_t)$ and is negligible compared with the first term in (16).

*2) Impact of the Non-i.i.d. Channels:* In general CoMP channels, the feedback overhead of $\mathrm{MS}_k$ depends on the large scale fading gains of its per-cell channels $\alpha_{k,b}$, $b = 1, \ldots, N_b$, as shown in $B_{k,\mathrm{sum}}^{\mathrm{CDI}}(2)$ and $B_{k,\mathrm{sum}}^{\mathrm{PA}}(2)$ in (11) and (12). In particular, since the geometric mean of a vector is not larger than its arithmetic mean, we have $\frac{(\prod_{l=1}^{N_b} \alpha_{k,l}^2)^{1/N_b}}{\frac{1}{N_b}\sum_{l=1}^{N_b} \alpha_{k,l}^2} \leq 1$ which implies $B_{k,\mathrm{sum}}^{\mathrm{CDI}}(2) \leq 0$, where the equality holds when $\alpha_{k,1}^2 = \cdots = \alpha_{k,N_b}^2$. This implies that the imbalance of the per-cell channel gains of $\mathrm{MS}_k$ can help reduce the required codebook size for the per-cell CDIs. Similarly, we have $(1 - \bar{\beta}_{k,b})\bar{\beta}_{k,b} \leq [(1 - \bar{\beta}_{k,b} + \bar{\beta}_{k,b})/2]^2 = 1/4$ which indicates $B_{k,\mathrm{sum}}^{\mathrm{PA}}(2) \leq -(N_b - 1)$, where the equality holds when $\bar{\beta}_{k,b} = 1/2$, $b = 1, \ldots, N_b$, $b \neq b_k$. In practical systems, the large scale fading gains of cross channels are not larger than that of local channel, i.e., $\alpha_{k,b}^2 \leq \alpha_{k,b_k}^2$, thus $\bar{\beta}_{k,b} = \frac{\alpha_{k,b}^2}{\sum_{l=1}^{N_b} \alpha_{k,l}^2} \leq \frac{\alpha_{k,b}^2}{\alpha_{k,b}^2 + \alpha_{k,b_k}^2} \leq 1/2$, $b \neq b_k$, where the equality of the last step holds when $\alpha_{k,b}^2 = \alpha_{k,b_k}^2$. This implies that the heterogeneous per-cell channel gains can also help reduce the number of bits for feeding back the PAs.

The numbers of bits for feeding back the per-cell CDIs of $\mathrm{MS}_k$ also depend on the large scale fading gains of its co-scheduled users, which is reflected in $\bar{\beta}_{j,b}$ in the term $B_{k,\mathrm{sum}}^{\mathrm{CDI}}(1)$ of (11), whose impact is weighted by $\gamma_k^{\mathrm{QoS}}$.

When the target SINR $\gamma_k^{\mathrm{QoS}}$ is low, the term $B_{k,\mathrm{sum}}^{\mathrm{CDI}}(1)$ can be approximated as

$$\begin{aligned} B_{k,\mathrm{sum}}^{\mathrm{CDI}}(1) \approx& (N_t - 1)\sum_{b=1}^{N_b} \frac{1}{N_t}\frac{1}{\mu_k}\gamma_k^{\mathrm{QoS}} \sum_{j=1, j\neq k}^{M} \bar{\beta}_{j,b} \\ =& \frac{N_t - 1}{N_t}\frac{1}{\mu_k}(M - 1)\gamma_k^{\mathrm{QoS}}, \end{aligned} \tag{17}$$

where the approximation is obtained by considering that $\log_2(1 + x) \approx x$ for small value of $x$, and the last step is derived from the fact that $\sum_{b=1}^{N_b} \bar{\beta}_{j,b} = 1$.

We can observe that in this case the feedback overhead is independent of the channel imbalance of the co-scheduled

users.

When the target SINR is high, the term can be approximated as

$$B_{k,\mathrm{sum}}^{\mathrm{CDI}}(1) \approx (N_t-1)\sum_{b=1}^{N_b}\log_2\left(\frac{\gamma_k^{\mathrm{QoS}}}{N_t\mu_k}\sum_{j=1,j\neq k}^{M}\bar{\beta}_{j,b}\right) \quad (18)$$

$$=N_b(N_t-1)\log_2\left(\frac{\gamma_k^{\mathrm{QoS}}}{N_t\mu_k}\right)+(N_t-1)\log_2\Gamma_k, \quad (19)$$

where $\Gamma_k=\prod_{b=1}^{N_b}(\sum_{j=1,j\neq k}^{M}\bar{\beta}_{j,b})$, and the approximation is from ignoring "1" in the $\log_2(\cdot)$ function and is accurate for large value of $\gamma_k^{\mathrm{QoS}}$.

Again, considering that the geometric mean of a vector is not larger than its arithmetic mean, we have $\Gamma_k \leq (1/N_b\sum_{b=1}^{N_b}\sum_{j=1,j\neq k}^{M}\bar{\beta}_{j,b})^{N_b}=[(M-1)/N_b]^{N_b}$, where the fact $\sum_{b=1}^{N_b}\bar{\beta}_{j,b}=1$ is applied and $\Gamma_k$ achieves its maximal value when $\sum_{j=1,j\neq k}^{M}\bar{\beta}_{j,1}=\cdots=\sum_{j=1,j\neq k}^{M}\bar{\beta}_{j,N_b}$. To gain useful insight, we consider two extreme scenarios. When only one user is co-scheduled with $\mathrm{MS}_k$, $\Gamma_k$ achieves its maximal value when $\bar{\beta}_{j,1}=\cdots=\bar{\beta}_{j,N_b}$, which is equivalent to $\alpha_{j,1}^2=\cdots=\alpha_{j,N_b}^2$. This implies that the channel imbalance of its co-scheduled user, $\mathrm{MS}_j$, helps reduce the feedback overhead of the desired user. When the number of co-scheduled users of $\mathrm{MS}_k$ tends to infinity and all the users are randomly distributed in the cooperative region, the value of $\sum_{j=1,j\neq k}^{M}\bar{\beta}_{j,b}=\sum_{j=1,j\neq k}^{M}\frac{\alpha_{j,b}^2}{\sum_{l=1}^{N_b}\alpha_{j,l}^2}$ tends to be a constant according to the law of large numbers. This suggests that when the number of users increases, the impact of the locations of the co-scheduled users will vanish. The numerical results in Section IV will show that the number of users do not need to be too large.

*3) Impact of the Target SINR and Cell-edge SNR:* From $B_{k,\mathrm{sum}}^{\mathrm{CDI}}(3)$ and $B_{k,\mathrm{sum}}^{\mathrm{PA}}(3)$ we see that the overhead is monotonically increasing with $\gamma_k^{\mathrm{QoS}}$ and decreasing with $\gamma_k^{\mathrm{Perf}}$. When $\gamma_k^{\mathrm{QoS}} \ll \gamma_k^{\mathrm{Perf}}$, the values of both $B_{k,\mathrm{sum}}^{\mathrm{CDI}}(3)$ and $B_{k,\mathrm{sum}}^{\mathrm{PA}}(3)$ approach 0, and the feedback overhead scales logarithmically with the target SINR. When the values of $\gamma_k^{\mathrm{QoS}}$ and $\gamma_k^{\mathrm{Perf}}$ are on the same order, the values of $B_{k,\mathrm{sum}}^{\mathrm{CDI}}(3)$ and $B_{k,\mathrm{sum}}^{\mathrm{PA}}(3)$ will dominate the total feedback overhead and their values will approach infinity rapidly as $\gamma_k^{\mathrm{QoS}}\to\gamma_k^{\mathrm{Perf}}$. This implies that the target SINR of a user should not be close to the maximal SINR with perfect CSI, otherwise its feedback overhead will soon become unacceptable.

The per-user feedback overhead depends on the cell-edge SNR, which is included in the expression of $\gamma_k^{\mathrm{Perf}}$ as shown in (10). When the cell-edge SNR increases, the value of $\gamma_k^{\mathrm{Perf}}$ will increase accordingly, and the feedback overhead of a user to support a given target SINR will reduce. This seems to be contrary to the analysis for single-cell systems [5], where the overhead should increase linearly with SNR. The inconsistence comes from the different design criteria. The required overhead in [5] is to ensure a constant rate loss, which corresponds to increase the QoS requirement when the cell-edge SNR grows according to our analysis in Section III.B.1), but what we analyzed is the minimal overhead to ensure a given QoS.

As will be shown later, allocating the total number of feedback bits to each user according to its required QoS can reduce the overall feedback overhead. Designing a feedback strategy to ensure a constant rate loss is spectrally efficient for downlink transmission, which however demands more uplink resource for the channel feedback [5]. By contrast, if we design a feedback strategy just to support the required QoS of each user, the overall feedback overhead will be reduced, which provides a good trade-off between the usage of the uplink and downlink resources.

## IV. Numerical and Simulation Results

In this section, we will verify our previous analytical results and analyze the required feedback overhead of CoMP system through numerical and simulation results.

The cell-edge SNR is set as 15 dB [21].[5] The codebooks used for quantizing the per-cell CDIs are generated by RVQ. In the simulation, the path-loss model is $\mathrm{PL}^{\mathrm{dB}}=35.3+37.6\log_{10}(d_{k,b})$, which is employed in 3GPP LTE, where $d_{k,b}$ (in meter) is the distance between $\mathrm{MS}_k$ and $\mathrm{BS}_b$. All simulation results are obtained by averaging over 5000 realizations of the small scale i.i.d. Rayleigh fading channels. We consider random user scheduling in the simulation unless otherwise specified, where the value of $\mu_k$ in (11) and (12) is $(N_bN_t-M+1)/(N_bN_t)$ [22]. When orthogonal scheduling is considered, $\mu_k=1$. In practice, $(N_bN_t-M+1)/(N_bN_t)<\mu_k<1$ since any well-designed scheduler will provide better average orthogonality than the random scheduler. Though our analytical results are applicable for general cases with arbitrary number of cooperative BSs, except for the last figure, we consider a simple but fundamental scenario with two cells, each has a group of users with the same number, and the two user-groups are symmetrically located at the cell-edge region, as shown in Fig. 1. That is to say, all users have identical distance with their local BSs. In this way we only need to show the performance of one user.

To show that our analysis using PUPC is also valid under practical power constraint, we consider PBPC in the simulations unless otherwise specified, where we employ the optimal power allocation that maximizes the sum rate under PBPC constraint [23].

All the simulation results in the sequel are obtained in the following way. We first find the total number of bits required by each user from (11) and (12) for a given $\gamma_k^{\mathrm{QoS}}$. Then we simulate the average SINR that can be achieved by the concerned feedback scheme with this bit number, such that all schemes for comparison have identical total number of bits for each user.

### A. Validation of the Analytical Analysis

Before we analyze the feedback overhead of the CoMP systems employing the minimal number of bits provided in *Proposition* 1, we first show the impact of the approximations introduced in the appendix, the impact of the assumption of PUPC and the approximations applied for the scaling law

[5]In [21], 20 dB was shown as a reasonable cell-edge SNR, which captures various settings in typical outdoor cellular networks. We set 5 dB lower since we take inter-cluster interference into account and treat it as AWGN.

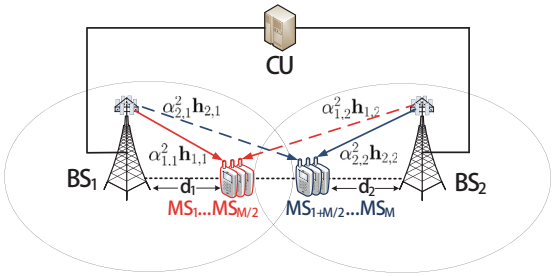


Fig. 1. An example of CoMP system, where the solid lines denote local channels while the dash lines denote cross channels of the users. The cell radius is 250 m. The users in the same cell are located in the same place, and the users are located on the line connecting the two BSs, such that the users with 0 dB and 6 dB channel energy imbalance are respectively located with 250 m and 205 m from their local BSs.

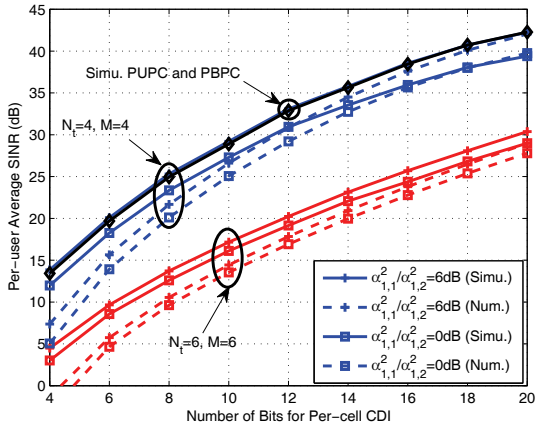


Fig. 2. Per-user average SINR versus the number of bits for the per-cell CDI. $N_b$ = 2. The bits for local CDI and cross CDI are identical, and the bits for PA are set as the half of those for per-cell CDI. The curves with legend 'Simu.' is the simulated SINR, and the curves with legend 'Num.' are numerically obtained from (A.20).

analysis. For concise, we evaluate the accuracy of the derived average SINR as well as the resulting per-user feedback overhead, instead of evaluating each of the approximations.

In Fig. 2 we compare the average SINR numerically obtained from (A.20) (where PUPC was assumed and several approximations were introduced) with the simulation results obtained by PBPC. To show where the gap between the numerical and simulation results comes from, we also provide the simulation result obtained under PUPC for the case where $\alpha_{1,1}^2/\alpha_{1,2}^2 = 6$ dB, $N_t = 4$ and $M = 4$. The results of other scenarios are similar, which are omitted for providing a clear figure. We can observe that the simulated result under the PUPC almost overlaps with that obtained with PBPC, which indicates that the gap is not caused by the PUPC assumption. Comparing the simulation results with the numerical results, we can see that the derived SINR is a lower bound of the simulated SINR, because we use a lower bound of the signal power in (A.12) and an upper bound of the interference power in (A.17), which become tight when the number of feedback bits increases.

In Fig. 3 we compare the per-user feedback overhead numerically obtained from (11) and (12) with the simulation results, from which we can further observe the impact of the

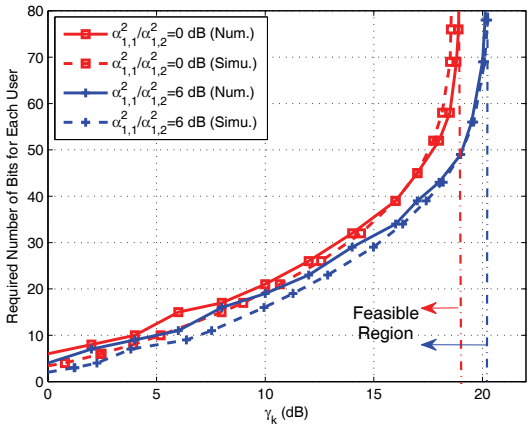


Fig. 3. Per-user feedback overhead versus the target SINR. $N_b$ = 2, $N_t$ = 4, $M$ = 4. The curves with legend 'Simu.' is the simulated results under PBPC, and the curves with legend 'Num.' are numerically obtained from (11) and (12) with the assumption of PUPC.

lower bound of the average SINR on the overhead analysis. We can see that the numerical results are close to the simulation results, and the numerical feedback overhead exceeds the simulated feedback overhead, which comes from the application of average SINR lower bound in (A.20). Furthermore, we can observe that for a given target SINR, the numbers of feedback bits of the users at different locations differ. For a given location of a user, the number of bits increases with the target SINR. The feedback overhead is more sensitive to the target QoS than the location.

In Fig. 4 we compare the numerical results of the term $B_{k,\text{sum}}^{\text{CDI}}(1)$ in (11), its approximations under low target SINR in (17) and high target SINR in (19). We can see that the approximated value in (17) is close to the true value under low target SINR (e.g., less than 8 dB), and the channel imbalance of a co-scheduled user has no impact on the results, which verifies our analysis. With the increase of the target SINR, the approximated value in (19) approaches the true value, and the channel imbalance of the co-scheduled user helps reduce the feedback overhead of the desired MS. This also agrees with our analysis.

### *B. Comparison with Other Feedback Schemes*

In Fig. 5 we compare the overhead of the optimal bit allocation based per-cell quantization scheme with a global quantization scheme, and several other per-cell codebook based schemes that employ independent codeword selection. The global codebook obtained from generalized Lloyd algorithm provides near-optimal performance, but the complexity of the codebook generation is prohibitive when the number of bits is large. For a fair comparison, we employ RVQ for the global quantization, which can be simulated easily using the statistics of its quantization error [5]. Because the available statistics of the quantization error with RVQ are obtained under the assumption of i.i.d. channels, we consider the "exact cell-edge" users, whose global channels are i.i.d. as well. In this case, the overhead using the global codebook in CoMP systems is the same as the overhead of an isolated single-cell MU-MIMO system with a $N_bN_t$-antenna BS.

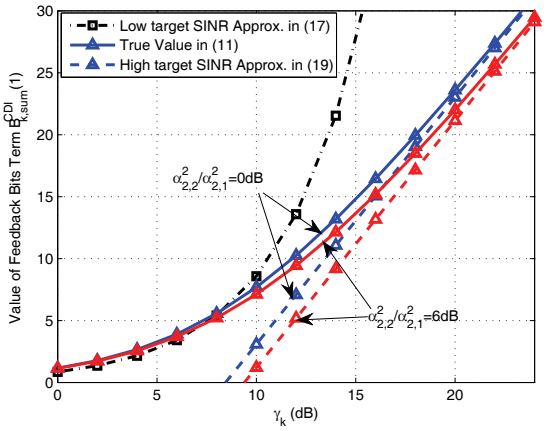


Fig. 4. Values of $B^{\mathrm{CDI}}_{k,\mathrm{sum}}(1)$ versus the target SINR of an "exact cell-edge user". $N_b$ = 2, $N_t$ = 4, $M$ = 2, where a co-scheduled user is with either 0 dB or 6 dB channel imbalance.

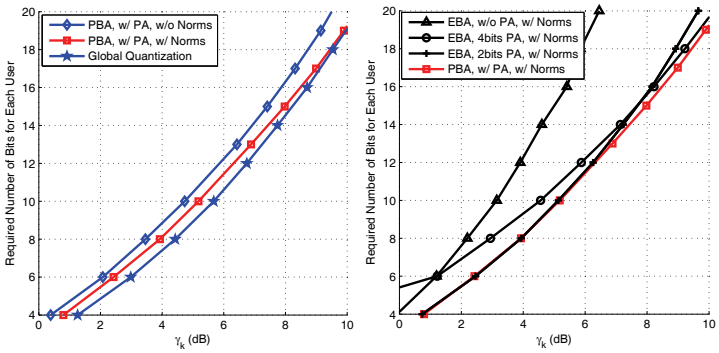


Fig. 5. Per-user feedback overhead versus the target SINR of the "exact cell-edge users". $N_b$ = 2, $N_t$ = 4, $M$ = 4. The legend 'PBA' denotes the result obtained from the optimal bit allocation in (9a) and (9b), and 'EBA' represents equal bit allocation among the per-cell CDIs. The legend 'w/ PA' represents the scheme where the PA difference is fed back, and those 'w/o PA', '2 bits PA' and '4 bits PA' respectively stand for the schemes with 0 bit, 2 bits and 4 bits for the PA feedback.

The simulation results for the feedback scheme with minimal overhead is obtained from the optimal bit allocation in (9a) and (9b) and with legend "PBA w/ PA and w/ norms", where the per-cell norms are assumed perfectly known at the CU whose overhead is not taken into account. The simulation results of all the other schemes are obtained with the same total number of bits for each user as the scheme with minimal overhead. To observe the impact of the per-cell channel norm feedback, we simulate a feedback scheme where the bit allocation is also optimal but the global CDI is re-constructed at the CU without using the per-cell norms, who is with legend "PBA w/ PA and w/o norms". We see that the performance loss led by not using per-cell norms is minor. By comparing this result with that of the global codebook where the per-cell channel norms have been implicitly quantized, we can also see how far the overhead of the structured codebook is from the global codebook. To observe the impact of the PA feedback, we provide simulation results of equal bit allocation between per-cell CDIs with 0 bit, 2 bits and 4 bits for PA feedback, where the per-cell channel norms are perfectly known at the CU. As expected, the performance is dramatically degraded without any PA feedback. These results verify the rationality to include the PA but exclude the per-cell channel norms in the feedback budget.

### *C. Impact of the User Location on the Total Feedback Bits and Bit Allocation*

It is widely recognized that cell-edge users should be served with CoMP to justify the extra cost and overhead. Nonetheless, these users are not necessarily the "exact cell-edge users" we defined.

In Fig. 6, we illustrate the impact of user location on the per-user total number of feedback bits and the bit allocation among per-cell CDIs and PA difference with numerical results, which are obtained from (11) and (12). Two user-groups are considered but they are no longer symmetrically located. We can observe that the increase of the channel imbalance of the desired user, $\mathrm{MS}_1$, leads to a reduction of its total number of feedback bits. When two BSs serve only two users, the channel

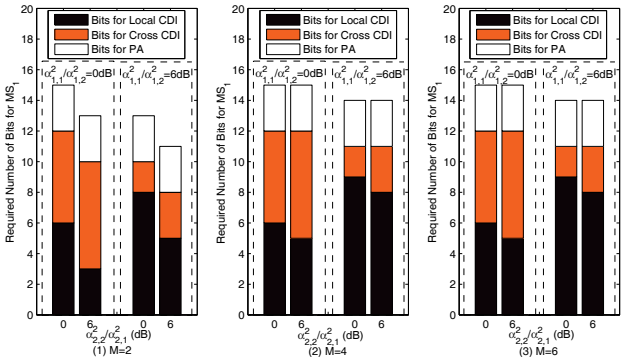


Fig. 6. Required total number of bits and the numbers of bits for the per-cell CDIs and PA for a user in cell 1, i.e., $\mathrm{MS}_1$. $N_b$ = 2, $N_t$ = 4. The increase of $\alpha^2_{2,2}/\alpha^2_{2,1}$ in $x$-axis represents the users in cell 2 moving from cell edge towards cell center. Similar representation holds for the channel energy imbalances of the users in cell 1, i.e., $\alpha^2_{1,1}/\alpha^2_{1,2}$. To show the impact of the active user number on the bit allocation between per-cell CDIs, we ensure the numbers of bits for different values of $M$ to be similar, then the target SINR of each user are respectively set as 12 dB, 6 dB and 2 dB for $M$ = 2, 4, 6.

imbalance of $\mathrm{MS}_2$ also leads to a feedback overhead reduction of $\mathrm{MS}_1$. When the number of users grows, the impact of the channel imbalance of the co-scheduled users of $\mathrm{MS}_1$ vanishes. This validates our former analysis.

### *D. Feedback Overhead of CoMP for Achieving $\eta$-fold of SINR of Non-CoMP*

Considering that $\mathrm{MS}_k$ is expected to achieve better performance under CoMP than Non-CoMP, here we set $\gamma_k^{\mathrm{QoS}} = \eta\gamma_k^{\mathrm{NC}}$, where $\gamma_k^{\mathrm{NC}}$ is the average SINR achieved under Non-CoMP transmission.

In Fig. 7, we show how many more feedback bits are required for CoMP to achieve $\eta$ times higher SINR than Non-CoMP, where the locations of users are given and the simulation setup is the same as that considered in Fig. 3. For Non-CoMP transmission, four bits are employed by each user to quantize its CDI. In the simulation, we first obtain $\gamma_k^{\mathrm{NC}}$ by averaging over small scale fading channels, then the target SINR of CoMP is set as $\gamma_k^{\mathrm{QoS}} = \eta\gamma_k^{\mathrm{NC}}$, with which we can

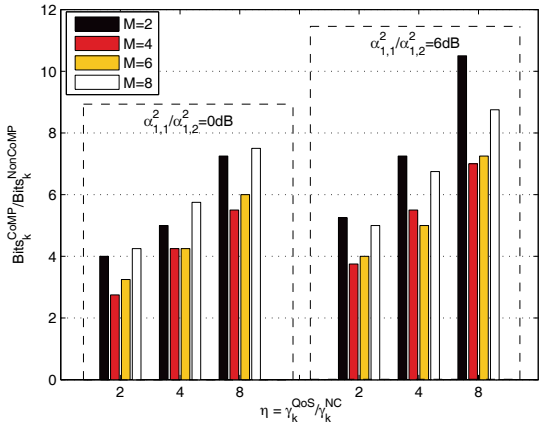


Fig. 7. Per-user feedback overhead ratio of CoMP over Non-CoMP versus $\eta$. $N_b$ = 2, $N_t$ = 4. With Non-CoMP transmission, four bits are applied for quantizing the CDI of each user, i.e., $\text{Bit}_k^{\text{NonCoMP}} = 4$.

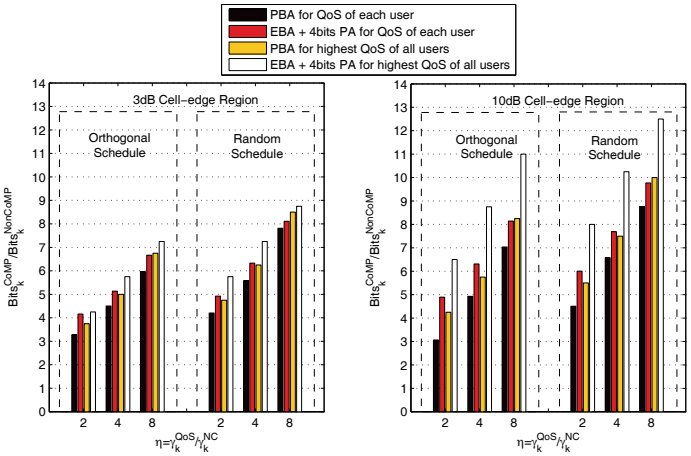


Fig. 8. Per-user feedback overhead ratio of CoMP over Non-CoMP versus $\eta$. $N_b$ = 3, $N_t$ = 4, $M$ = 6 and $\text{Bit}_k^{\text{NonCoMP}} = 4$. The legends "PBA for QoS of each user" and "PBA for highest QoS of all users" represent the simulation results obtained from the optimal bit allocation in (9a) and (9b) according to the required QoS of each user and according to the highest QoS of all users. The legends "EBA + 4bits PA for QoS of each user" and "EBA + 4bits PA for highest QoS of all users" denote the simulation results obtained from equal bit allocation among per-cell CDIs and 4 bits for PA according to the QoS of each user and according to the highest QoS, respectively. To ensure the required QoS of $\text{MS}_k$ with equal bit allocation, the bit numbers for its per-cell CDIs are obtained from the optimization problem shown in (8) after introducing extra constraints $B_{k,1}^{\text{CDI}} = \cdots = B_{k,N_b}^{\text{CDI}}$ and $B_{k,b}^{\text{PA}} = 4$.

numerically obtain the total number of feedback bits from (11) and (12). We can observe that the user located at the "exact cell edge" requires minimal feedback bits for CoMP to achieve $\eta$-fold of the SINR over Non-CoMP. This seems to be inconsistent with the previous results shown in Fig. 3, which comes from the difference in the QoS requirement. When a user is located at the "exact cell edge", its SINR under Non-CoMP is the lowest, thereby it is easy for CoMP to achieve higher SINR than Non-CoMP. We can also observe that the per-user feedback overhead when the number of scheduled users, $M$, is 2 or 8 is larger than that in the cases when $M$ is 4 or 6. This is because when $M = 2$, each cell has one user, then there is no MUI in Non-CoMP and $\gamma_k^{\text{NC}}$ is high. As a result, more feedback bits are required for CoMP to achieve the SINR enhancement. On the other hand, the overhead of CoMP increases with the number of co-scheduled users $M$ as we analyzed in Section III.B.1). Consequently, more bits are required to be fed back for the case $M = 8$.

Finally, in Fig. 8 we provide simulation results in a more realistic setting. Three faced sectors form a cooperative cluster. The sector antenna power gain is a function of the horizontal angle $\phi$ (in degrees) follows LTE specification [24], i.e., $\text{AG}^{\text{dB}} = 14 - \min\{12(\frac{\phi}{70})^2, 20\}$, $-\pi < \phi < \pi$. The users are randomly distributed in a "cell-edge region", where for any $\text{MS}_k$ $\max\limits_{l=1,\ldots,N_b} \frac{\alpha^2_{k,b_k}}{\alpha^2_{k,l}}$ is less than a predefined value. For each randomly distributed user, its feedback overhead is obtained numerically from (11) and (12). The results are obtained by averaging over 50 random locations of the users. To show the impact of the user scheduling, we consider both random and orthogonal scheduling, which are obtained by setting $\mu_k = (N_bN_t - M + 1)/(N_bN_t)$ and $\mu_k = 1$, respectively. To show the benefit of the bit allocation among the users according to their required QoS and the bit allocation among the per-cell CDIs according to the user location, we provide the results of two conservative feedback schemes. One scheme uses the optimal bit allocation for each user by setting the target SINR as the highest $\gamma^{\text{QoS}}$ among all users. The other employs equal bit allocation for all the users, and the number of bits is obtained for supporting the highest $\gamma^{\text{QoS}}$ of all users. The results show that to achieve the minimal feedback overhead of CoMP systems, it is critical to allocate the number of bits among different users and different per-cell CDIs. Allocating the total number of bits for each user according to its required QoS can significantly reduce the feedback overhead relative to the conservative bit allocation.

## V. Conclusions

In this paper, we analyzed the minimal feedback overhead required for quantizing the channel direction information to facilitate downlink precoding of coherent CoMP transmission in order to ensure the required performance of each user. We considered per-cell codebook based feedback strategy, which is scalable and backward compatible, where a phase ambiguity information is fed back and a low complexity independent codeword selection is applied. Our analysis showed that to support a target performance, the feedback overhead of each user should increase with the overall number of antennas at the cooperative BSs and increase with the decrease of the cell-edge SNR. Through simulations, we validated the analysis and showed how much more feedback overhead is required for the CoMP system to provide a given performance gain over the Non-CoMP system. Analytical and simulation results showed that to reduce the overall feedback overhead of the cooperative network, the bit allocation among multiple users is critical by exploiting the difference in their performance gain, and the bit allocation among per-cell codebooks is beneficial by exploiting the imbalance of the average channel gains. In practice, the feedback bit number can be allocated semi-dynamically according to the number of cooperative BSs, the number of antennas at each BS, the cell-edge SNR, the target average SINR of each user, as well as the user location.

## APPENDIX A
## PROOF OF PROPOSITION 1

We first show that the optimization problem that minimizes the overhead of $\text{MS}_k$ to achieve its target average SINR can be formulated as the problem shown in (8). To this end, we will derive the average SINR constraint in (8b). Then, we show that the closed-form solution of the problem is (9).

### A. Derivation of the Average SINR Constraint in (8b)

To obtain the per-user average SINR constraint, we will first derive the expressions of the estimated instantaneous received signal power $\hat{\text{S}}_k$ and interference power $\hat{\text{I}}_k$ of $\text{MS}_k$, with which the estimated instantaneous SINR can be obtained as $\hat{\gamma}_k = \frac{\hat{\text{S}}_k}{\hat{\text{I}}_k+\sigma_k^2}$. Then, we strive to derive the expression of the average SINR $\mathbb{E}\{\hat{\gamma}_k\}$, from which constraint (8b) can be derived. In order to obtain an explicit expression of the average SINR for the tractability of the optimization, we introduce some approximations in the following derivations.

Denote the angle between the ZFBF vector $\boldsymbol{v}_k$ and the quantization vector of global CDI $\hat{\boldsymbol{g}}_k$ as $\psi_k$. Then $\cos^2\psi_k = \left|\boldsymbol{v}_k^H \frac{\hat{\boldsymbol{g}}_k}{\|\hat{\boldsymbol{g}}_k\|}\right|^2$ and $\boldsymbol{v}_k$ can be expressed as $\boldsymbol{v}_k = \cos\psi_k \frac{\hat{\boldsymbol{g}}_k}{\|\hat{\boldsymbol{g}}_k\|} + \sin\psi_k \boldsymbol{e}_k$ by applying orthogonal expansion, where $\frac{\hat{\boldsymbol{g}}_k}{\|\hat{\boldsymbol{g}}_k\|}$ and $\boldsymbol{e}_k \in \mathbb{C}^{N_b N_t \times 1}$ are the two orthogonal basis, and $\boldsymbol{e}_k$ is a unit-norm vector distributed in the $(N_b N_t - 1)$-dimensional null space of $\hat{\boldsymbol{g}}_k$.

From (7) we know that when $j = k$, $\hat{q}_{k,k}^{\text{MMSE}}$ is the MMSE estimate of the signal power of $\text{MS}_k$, i.e., $\hat{\text{S}}_k = \hat{q}_{k,k}^{\text{MMSE}}$. Then the MMSE estimate of the signal power of $\text{MS}_k$ can be derived from (7) as

$$\begin{aligned}\hat{\text{S}}_k =& \hat{q}_{k,k}^{\text{MMSE}} \\ =& \frac{N_b P_0}{M}\mathbb{E}_{\boldsymbol{D}_k}\{|\hat{\boldsymbol{g}}_k^H \boldsymbol{D}_k^H \boldsymbol{v}_k|^2\} + \frac{N_b P_0}{M}\mathbb{E}_{\boldsymbol{s}_{\boldsymbol{g}_k}}\{|\boldsymbol{s}_{\boldsymbol{g}_k}^H \boldsymbol{v}_k|^2\} \qquad \text{(A.1)}\\ \overset{(a)}{=}& \frac{N_b P_0}{M}\mathbb{E}_{\boldsymbol{D}_k}\left\{\left|\hat{\boldsymbol{g}}_k^H \boldsymbol{D}_k^H \left(\cos\psi_k \frac{\hat{\boldsymbol{g}}_k}{\|\hat{\boldsymbol{g}}_k\|} + \sin\psi_k \boldsymbol{e}_k\right)\right|^2\right\} \\ &+ \frac{N_b P_0}{M}\mathbb{E}_{\boldsymbol{s}_{\boldsymbol{g}_k}}\left\{\left|\boldsymbol{s}_{\boldsymbol{g}_k}^H \left(\cos\psi_k \frac{\hat{\boldsymbol{g}}_k}{\|\hat{\boldsymbol{g}}_k\|} + \sin\psi_k \boldsymbol{e}_k\right)\right|^2\right\} \qquad \text{(A.2)}\\ \overset{(b)}{\approx}& \cos^2\psi_k \frac{N_b P_0}{M} \underbrace{\mathbb{E}_{\boldsymbol{D}_k}\left\{\left|\frac{\hat{\boldsymbol{g}}_k^H}{\|\hat{\boldsymbol{g}}_k\|}\boldsymbol{D}_k^H \hat{\boldsymbol{g}}_k\right|^2\right\}}_{\hat{\text{S}}_k(1)} \\ &+ \cos^2\psi_k \frac{N_b P_0}{M} \underbrace{\mathbb{E}_{\boldsymbol{s}_{\boldsymbol{g}_k}}\left\{\left|\boldsymbol{s}_{\boldsymbol{g}_k}^H \frac{\hat{\boldsymbol{g}}_k}{\|\hat{\boldsymbol{g}}_k\|}\right|^2\right\}}_{\hat{\text{S}}_k(2)}, \qquad \text{(A.3)}\end{aligned}$$

where $(a)$ is obtained by substituting the expression $\boldsymbol{v}_k = \cos\psi_k \frac{\hat{\boldsymbol{g}}_k}{\|\hat{\boldsymbol{g}}_k\|} + \sin\psi_k \boldsymbol{e}_k$, and $(b)$ comes by approximating the direction of beamforming vector as the quantized global CDI, i.e., $\sin\psi_k \approx 0$, under which the term $\sin\psi_k \boldsymbol{e}_k$ in (A.2) can be ignored. This approximation is accurate when the quantized global CDI of $\text{MS}_k$ and that of its co-scheduled users are nearly orthogonal.

By substituting the expression of $\hat{\boldsymbol{g}}_k$ and $\boldsymbol{D}_k$ shown in (4) and (6) into the expression of $\hat{\text{S}}_k(1)$ in the right-hand side of (A.3), we can obtain

$$\begin{aligned}\hat{\text{S}}_k(1) =& \mathbb{E}_{\Delta\omega_{k,b},\theta_{k,b}}\left\{\frac{1}{\|\hat{\boldsymbol{g}}_k\|^2}\Big|\alpha_{k,b_k}^2\|\boldsymbol{h}_{k,b_k}\|^2\cos\theta_{k,b_k} \right. \\ &\left. + \sum\nolimits_{b=1,b\neq b_k}^{N_b} \alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2\cos\theta_{k,b} e^{j\Delta\omega_{k,b}}\Big|^2\right\}. \qquad \text{(A.4)}\end{aligned}$$

To simplify the notations and for the sake of clarity, we define $\boldsymbol{u}_k \triangleq [u_{k,1},\ldots,u_{k,N_b}]^H$, with $u_{k,b} = \frac{\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2\cos\theta_{k,b}}{\|\hat{\boldsymbol{g}}_k\|} = \frac{\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2\cos\theta_{k,b}}{(\sum_{l=1}^{N_b}\alpha_{k,l}^2\|\boldsymbol{h}_{k,l}\|^2)^{1/2}}$, and define $\Delta\boldsymbol{\omega}_k \triangleq -[\Delta\omega_{k,1},\ldots,\Delta\omega_{k,b_k-1},0,\Delta\omega_{k,b_k+1},\ldots,\Delta\omega_{k,N_b}]^H$, whose elements are the quantization errors of PA differences. Then we can further express $\hat{\text{S}}_k(1)$ as

$$\begin{aligned}\hat{\text{S}}_k(1) =& \mathbb{E}_{\Delta\omega_{k,b},\theta_{k,b}}\left\{\left|(e^{j\Delta\boldsymbol{\omega}_k})^H\boldsymbol{u}_k\right|^2\right\} \\ =& \mathbb{E}_{\Delta\omega_{k,b},\theta_{k,b}}\left\{(e^{j\Delta\boldsymbol{\omega}_k})^H\boldsymbol{u}_k\boldsymbol{u}_k^H e^{j\Delta\boldsymbol{\omega}_k}\right\} \qquad \text{(A.5)}\\ \overset{(a)}{\approx}& \mathbb{E}_{\Delta\omega_{k,b},\theta_{k,b}}\left\{\Delta\boldsymbol{\omega}_k^H\left[\boldsymbol{u}_k\boldsymbol{u}_k^H - \text{diag}\left\{\boldsymbol{u}_k\boldsymbol{u}_k^H\mathbf{1}_{N_b}\right\}\right]\Delta\boldsymbol{\omega}_k\right\} \\ &+ \mathbb{E}_{\theta_{k,b}}\{|\boldsymbol{u}_k^H\mathbf{1}_{N_b}|^2\} \qquad \text{(A.6)}\\ \overset{(b)}{=}& \sum\nolimits_{b=1,b\neq b_k}^{N_b}\mathbb{E}_{\Delta\omega_{k,b},\theta_{k,b}}\left\{|\Delta\omega_{k,b}|^2u_{k,b}^2\right\} \\ &- \sum\nolimits_{b=1,b\neq b_k}^{N_b}\mathbb{E}_{\Delta\omega_{k,b},\theta_{k,b}}\left\{|\Delta\omega_{k,b}|^2u_{k,b}\boldsymbol{u}_k^H\mathbf{1}_{N_b}\right\} \\ &+ \sum\nolimits_{b=1}^{N_b}\sum\nolimits_{a=1}^{N_b}\mathbb{E}_{\theta_{k,b},\theta_{k,a}}\{u_{k,b}u_{k,a}\} \qquad \text{(A.7)}\\ =& -\sum_{b=1,b\neq b_k}^{N_b}\sum_{a=1,a\neq b}^{N_b}\mathbb{E}\{|\Delta\omega_{k,b}|^2\}\mathbb{E}_{\theta_{k,b},\theta_{k,a}}\{u_{k,b}u_{k,a}\} \\ &+ \sum\nolimits_{b=1}^{N_b}\sum\nolimits_{a=1}^{N_b}\mathbb{E}_{\theta_{k,b},\theta_{k,a}}\{u_{k,b}u_{k,a}\}, \qquad \text{(A.8)}\end{aligned}$$

where $(a)$ is obtained by approximating (A.5) as its second order Taylor expansion. Similar approximation was applied in Section VII of [25] with detailed derivation, which is omitted here for brevity. $(b)$ comes by considering $\mathbb{E}\{\Delta\omega_{k,b}\} = 0$. $\mathbf{1}_{N_b}$ denotes the column vector of size $N_b$ with all elements as 1.

The term $\mathbb{E}_{\theta_{k,b},\theta_{k,a}}\{u_{k,b}u_{k,a}\}$ in (A.8) can be derived as

$$\begin{aligned}\xi_{k,b,a} \triangleq& \mathbb{E}_{\theta_{k,b},\theta_{k,a}}\{u_{k,b}u_{k,a}\} \\ \overset{(a)}{\approx}& \beta_{k,b}\alpha_{k,a}^2\|\boldsymbol{h}_{k,a}\|^2\mathbb{E}_{\theta_{k,b},\theta_{k,a}}\left\{(1-\frac{1}{2}\sin^2\theta_{k,b})(1-\frac{1}{2}\sin^2\theta_{k,a})\right\} \\ \overset{(b)}{\approx}& \beta_{k,b}\alpha_{k,a}^2\|\boldsymbol{h}_{k,a}\|^2\left(1-\frac{1}{2}2^{-\frac{B_{k,b}^{\text{CDI}}}{N_t-1}}-\frac{1}{2}2^{-\frac{B_{k,a}^{\text{CDI}}}{N_t-1}}\right), \qquad \text{(A.9)}\end{aligned}$$

where $\beta_{k,b} = \frac{\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2}{\sum_{l=1}^{N_b}\alpha_{k,l}^2\|\boldsymbol{h}_{k,l}\|^2}$, $(a)$ comes from approximating $\cos\theta_{k,b} = (1-\sin^2\theta_{k,b})^{\frac{1}{2}}$ as its first order Taylor expansion, and $(b)$ is obtained by omitting the high order term $\frac{1}{4}\sin^2\theta_{k,b}\sin^2\theta_{k,a}$ and approximating $\mathbb{E}\{\sin^2\theta_{k,b}\}$ by its upper bound with RVQ [5]. These approximations are accurate when the codebook sizes are large.

Considering uniform scalar quantization for PA difference and substituting (A.9) into (A.8) we have

$$\begin{aligned}\hat{\mathrm{S}}_k(1) &\overset{(a)}{>} \sum_{b=1}^{N_b}\sum_{a=1}^{N_b}\beta_{k,b}\alpha_{k,a}^2\|\boldsymbol{h}_{k,a}\|^2\Big(1-\frac{1}{2}2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t-1}}-\frac{1}{2}2^{-\frac{B_{k,a}^{\mathrm{CDI}}}{N_t-1}}\Big)\\ &\quad-\sum_{b=1,b\neq b_k}^{N_b}\frac{\pi^2}{3}2^{-2B_{k,b}^{\mathrm{PA}}}\sum_{a=1,a\neq b}^{N_b}\beta_{k,b}\alpha_{k,a}^2\|\boldsymbol{h}_{k,a}\|^2\\ &\overset{(b)}{=}\sum_{a=1}^{N_b}\alpha_{k,a}^2\|\boldsymbol{h}_{k,a}\|^2-\frac{1}{2}\sum_{b=1}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2 2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t-1}}\\ &\quad-\frac{1}{2}\sum_{a=1}^{N_b}\alpha_{k,a}^2\|\boldsymbol{h}_{k,a}\|^2 2^{-\frac{B_{k,a}^{\mathrm{CDI}}}{N_t-1}}\\ &\quad-\frac{\pi^2}{3}\sum_{b=1,b\neq b_k}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2(1-\beta_{k,b})2^{-2B_{k,b}^{\mathrm{PA}}}\\ &=\sum_{b=1}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2\Big(1-2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t-1}}\Big)\\ &\quad-\frac{\pi^2}{3}\sum_{b=1,b\neq b_k}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2(1-\beta_{k,b})2^{-2B_{k,b}^{\mathrm{PA}}},\end{aligned}\tag{A.10}$$

where $(a)$ is because $\xi_{k,b,a} < \beta_{k,b}\alpha_{k,a}^2\|\boldsymbol{h}_{k,a}\|^2$, and $(b)$ is derived from two facts: 1) $\sum_{b=1}^{N_b}\beta_{k,b}=1$, and 2) $\sum_{a=1,a\neq b}^{N_b}\beta_{k,b}\alpha_{k,a}^2\|\boldsymbol{h}_{k,a}\|^2 = \frac{\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2(\sum_{l=1}^{N_b}\alpha_{k,l}^2\|\boldsymbol{h}_{k,l}\|^2-\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2)}{\sum_{l=1}^{N_b}\alpha_{k,l}^2\|\boldsymbol{h}_{k,l}\|^2} = \alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2(1-\beta_{k,b})$.

After substituting the expressions of $\hat{\boldsymbol{g}}_k$ and $\boldsymbol{s}_{\boldsymbol{g}_k}$ shown in (4) and (6), and considering the fact that the per-cell CDI quantization error $\boldsymbol{s}_{k,b}$ and the quantized per-cell CDI $\hat{\boldsymbol{h}}_{k,b}$ are mutually orthogonal, the expression of $\hat{\mathrm{S}}_k(2)$ in the right hand side of (A.3) can be derived as

$$\begin{aligned}\hat{\mathrm{S}}_k(2)&=\mathbb{E}_{\boldsymbol{s}_{k,b},\theta_{k,b}}\left\{\left|\sum_{b=1}^{N_b}\frac{\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2\sin\theta_{k,b}e^{j\hat{\omega}_{k,b}}}{(\sum_{l=1}^{N_b}\alpha_{k,l}^2\|\boldsymbol{h}_{k,l}\|^2)^{1/2}}\boldsymbol{s}_{k,b}^H\hat{\boldsymbol{h}}_{k,b}\right|^2\right\}\\ &=0.\end{aligned}\tag{A.11}$$

With (A.3), (A.10) and (A.11), we can obtain a lower bound of the MMSE estimate of the signal power for $\mathrm{MS}_k$ as

$$\begin{aligned}\hat{\mathrm{S}}_k^{\mathrm{LB}}&=\cos^2\psi_k\frac{N_bP_0}{M}\Bigg[\sum_{b=1}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2(1-2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t-1}})\\ &\quad-\frac{\pi^2}{3}\sum_{b=1,b\neq b_k}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2(1-\beta_{k,b})2^{-2B_{k,b}^{\mathrm{PA}}}\Bigg].\end{aligned}\tag{A.12}$$

The MMSE estimate of the interference power of $\mathrm{MS}_k$ can be obtained from (7) as

$$\begin{aligned}\hat{\mathrm{I}}_k&=\sum_{j=1,j\neq k}^{M}\hat{q}_{k,j}^{\mathrm{MMSE}}=\frac{N_bP_0}{M}\underbrace{\sum_{j=1,j\neq k}^{M}\mathbb{E}_{\boldsymbol{D}_k}\{|\hat{\boldsymbol{g}}_k^H\boldsymbol{D}_k^H\boldsymbol{v}_j|^2\}}_{\hat{\mathrm{I}}_k(1)}\\ &\quad+\frac{N_bP_0}{M}\underbrace{\sum_{j=1,j\neq k}^{M}\mathbb{E}_{\boldsymbol{s}_{\boldsymbol{g}_k}}\{|\boldsymbol{s}_{\boldsymbol{g}_k}^H\boldsymbol{v}_j|^2\}}_{\hat{\mathrm{I}}_k(2)}.\end{aligned}\tag{A.13}$$

The closed-form expressions of $\hat{\mathrm{I}}_k(1)$ and $\hat{\mathrm{I}}_k(2)$ are hard to derive. Instead, we derive their upper bounds respectively in the following.

The upper bound of term $\hat{\mathrm{I}}_k(1)$ can be derived as

$$\begin{aligned}\hat{\mathrm{I}}_k(1)&\overset{(a)}{=}\sum\nolimits_{j=1,j\neq k}^{M}\mathbb{E}_{\boldsymbol{D}_k}\{|\hat{\boldsymbol{g}}_k^H\boldsymbol{D}_k^H\boldsymbol{P}_{\hat{\boldsymbol{g}}_k}^{\perp}\boldsymbol{v}_j|^2\}\\ &\overset{(b)}{=}\mathbb{E}_{\boldsymbol{D}_k}\Big\{\mathrm{tr}\Big\{\boldsymbol{P}_{\hat{\boldsymbol{g}}_k}^{\perp}\boldsymbol{D}_k\hat{\boldsymbol{g}}_k\hat{\boldsymbol{g}}_k^H\boldsymbol{D}_k^H\boldsymbol{P}_{\hat{\boldsymbol{g}}_k}^{\perp}\sum\nolimits_{j=1,j\neq k}^{M}\boldsymbol{v}_j\boldsymbol{v}_j^H\Big\}\Big\}\\ &\overset{(c)}{\leq}\lambda_k^{\max}\mathbb{E}_{\boldsymbol{D}_k}\Big\{\mathrm{tr}\{\boldsymbol{P}_{\hat{\boldsymbol{g}}_k}^{\perp}\boldsymbol{D}_k\hat{\boldsymbol{g}}_k\hat{\boldsymbol{g}}_k^H\boldsymbol{D}_k^H\boldsymbol{P}_{\hat{\boldsymbol{g}}_k}^{\perp}\}\Big\}\\ &\overset{(d)}{=}\lambda_k^{\max}\mathbb{E}_{\boldsymbol{D}_k}\Big\{\hat{\boldsymbol{g}}_k^H\boldsymbol{D}_k^H\boldsymbol{P}_{\hat{\boldsymbol{g}}_k}^{\perp}\boldsymbol{D}_k\hat{\boldsymbol{g}}_k\Big\}\\ &=\lambda_k^{\max}\mathbb{E}_{\boldsymbol{D}_k}\{\hat{\boldsymbol{g}}_k^H\boldsymbol{D}_k^H\boldsymbol{D}_k\hat{\boldsymbol{g}}_k\}-\lambda_k^{\max}\mathbb{E}_{\boldsymbol{D}_k}\Big\{\Big|\frac{\hat{\boldsymbol{g}}_k^H}{\|\hat{\boldsymbol{g}}_k\|}\boldsymbol{D}_k^H\hat{\boldsymbol{g}}_k\Big|^2\Big\}\\ &\overset{(e)}{=}\lambda_k^{\max}\sum_{b=1}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2(1-2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t-1}})-\lambda_k^{\max}\hat{\mathrm{S}}_k(1)\\ &<\lambda_k^{\max}\frac{\pi^2}{3}\sum_{b=1,b\neq b_k}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2\,(1-\beta_{k,b})\,2^{-2B_{k,b}^{\mathrm{PA}}},\end{aligned}\tag{A.14}$$

where $\lambda_k^{\max}$ is the largest eigenvalue of matrix $\sum_{j=1,j\neq k}^{M}\boldsymbol{v}_j\boldsymbol{v}_j^H$, $(a)$ is because $\hat{\boldsymbol{g}}_k^H\boldsymbol{v}_j=0$ when ZFBF is applied, then by defining $\boldsymbol{P}_{\hat{\boldsymbol{g}}_k}^{\perp}=\mathbf{I}_{N_bN_t}-\frac{\hat{\boldsymbol{g}}_k\hat{\boldsymbol{g}}_k^H}{\hat{\boldsymbol{g}}_k^H\hat{\boldsymbol{g}}_k}$ as the orthogonal projection matrix of $\hat{\boldsymbol{g}}_k$, $\boldsymbol{v}_j$ can be expressed as $\boldsymbol{v}_j=\boldsymbol{P}_{\hat{\boldsymbol{g}}_k}^{\perp}\boldsymbol{v}_j$, $(b)$ is because $\mathrm{tr}\{\boldsymbol{AB}\}=\mathrm{tr}\{\boldsymbol{BA}\}$, and $\mathrm{tr}\{\cdot\}$ represents the trace of matrix, $(c)$ is obtained by the fact that $\mathrm{tr}\{\boldsymbol{AB}\}\leq\lambda_{\boldsymbol{B}}^{\max}\mathrm{tr}\{\boldsymbol{A}\}$ for matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and $\lambda_{\boldsymbol{B}}^{\max}$ is the largest eigenvalue of matrix $\boldsymbol{B}$, $(d)$ is owing to the fact that $(\boldsymbol{P}_{\hat{\boldsymbol{g}}_k}^{\perp})^2=\boldsymbol{P}_{\hat{\boldsymbol{g}}_k}^{\perp}$, $(e)$ comes from substituting the expressions of $\hat{\boldsymbol{g}}_k$ and $\boldsymbol{D}_k$ shown in (4) and (6) and approximating $\mathbb{E}\{\sin^2\theta_{k,b}\}$ by its upper bound under RVQ, and the last inequality comes after substituting the lower bound of $\hat{\mathrm{S}}_k(1)$ in (A.10).

The upper bound of term $\hat{\mathrm{I}}_k(2)$ in (A.13) can be derived as

$$\begin{aligned}\hat{\mathrm{I}}_k(2)&=\sum\nolimits_{j=1,j\neq k}^{M}\boldsymbol{v}_j^H\mathbb{E}\{\boldsymbol{s}_{\boldsymbol{g}_k}\boldsymbol{s}_{\boldsymbol{g}_k}^H\}\boldsymbol{v}_j\\ &\overset{(a)}{=}\frac{1}{N_t}\sum_{j=1,j\neq k}^{M}\sum_{b=1}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2\|\boldsymbol{v}_{j,b}\|^2\mathbb{E}\{\sin^2\theta_{k,b}\}\end{aligned}\tag{A.15}$$

$$\begin{aligned}&\overset{(b)}{\approx}\frac{1}{N_t}\sum_{j=1,j\neq k}^{M}\sum_{b=1}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2\frac{\alpha_{j,b}^2\|\boldsymbol{h}_{j,b}\|^2\cos^2\psi_j}{\sum_{l=1}^{N_b}\alpha_{j,l}^2\|\boldsymbol{h}_{j,l}\|^2}2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t-1}}\\ &\overset{(c)}{\leq}\frac{1}{N_t}\sum_{j=1,j\neq k}^{M}\sum_{b=1}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2\beta_{j,b}2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t-1}},\end{aligned}\tag{A.16}$$

where $(a)$ comes by considering the facts that $\mathbb{E}\{\boldsymbol{s}_{k,b}\boldsymbol{s}_{k,b}^H\}=\frac{1}{N_t}\mathbf{I}_{N_t}$ and $\mathbb{E}\{\boldsymbol{s}_{k,b}\boldsymbol{s}_{k,a}^H\}=\mathbf{0}_{N_t}$ for $b\neq a$, then we can obtain $\mathbb{E}\{\boldsymbol{s}_{\boldsymbol{g}_k}\boldsymbol{s}_{\boldsymbol{g}_k}^H\}=\frac{1}{N_t}\mathrm{diag}\Big\{\alpha_{k,1}^2\|\boldsymbol{h}_{k,1}\|^2\mathbb{E}\{\sin^2\theta_{k,1}\}\mathbf{I}_{N_t},\ldots,\alpha_{k,N_b}^2\|\boldsymbol{h}_{k,N_b}\|^2\mathbb{E}\{\sin^2\theta_{k,N_b}\}\mathbf{I}_{N_t}\Big\}$, $(b)$ is obtained from the approximation $\sin\psi_k\approx 0$, which is accurate when the quantized global CDI of $\mathrm{MS}_k$ and that of its co-scheduled users are nearly orthogonal, and $(c)$ is because $\cos^2\psi_j\leq 1$.

With (A.13), (A.14) and (A.16), we can obtain an upper bound of the MMSE estimate of interference power for $\mathrm{MS}_k$

as

$$\hat{\mathrm{I}}_k^{\mathrm{UB}} = \frac{N_b P_0}{N_t M}\Bigg[\sum_{j=1,j\neq k}^{M}\sum_{b=1}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2\beta_{j,b}2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t-1}} + N_t\lambda_k^{\max}\frac{\pi^2}{3}\sum_{b=1,b\neq b_k}^{N_b}\alpha_{k,b}^2\|\boldsymbol{h}_{k,b}\|^2(1-\beta_{k,b})2^{-2B_{k,b}^{\mathrm{PA}}}\Bigg]. \tag{A.17}$$

To derive a closed-form expression of the average SINR $\mathbb{E}\{\hat{\gamma}_k\}$ for a tractable optimization, we consider the asymptotic regime with large $N_t$. According to the law of large numbers, we have $\lim_{N_t\to\infty}\|\boldsymbol{h}_{k,b}\|^2 = N_t$. Then, $\lim_{N_t\to\infty}\beta_{j,b} = \frac{\alpha_{j,b}^2}{\sum_{l=1}^{N_b}\alpha_{j,l}^2} \triangleq \bar{\beta}_{j,b}$ and $\lim_{N_t\to\infty}\|\boldsymbol{h}_{k,b}\|^2\beta_{k,b} = N_t\frac{\alpha_{k,b}^2}{\sum_{l=1}^{N_b}\alpha_{k,l}^2} = N_t\bar{\beta}_{k,b}$. Similar to the proof in *Proposition 8* in [26], we can show that $\lim_{N_t\to\infty}\lambda_k^{\max} = 1$, which is omitted here for brevity. This implies that $\lim_{N_t\to\infty}\cos^2\psi_k = 1$. Then the signal power and interference power in (A.12) and (A.17) become

$$\hat{\mathrm{S}}_{k,\lim}^{\mathrm{LB}} \triangleq \lim_{N_t\to\infty}\hat{\mathrm{S}}_k^{\mathrm{LB}} = \frac{N_b P_0}{M}N_t\Bigg[\sum\nolimits_{b=1}^{N_b}\alpha_{k,b}^2(1-2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t-1}}) - \frac{\pi^2}{3}\sum\nolimits_{b=1,b\neq b_k}^{N_b}\alpha_{k,b}^2\left(1-\bar{\beta}_{k,b}\right)2^{-2B_{k,b}^{\mathrm{PA}}}\Bigg], \tag{A.18}$$

$$\hat{\mathrm{I}}_{k,\lim}^{\mathrm{UB}} \triangleq \lim_{N_t\to\infty}\hat{\mathrm{I}}_k^{\mathrm{UB}} = \frac{N_b P_0}{M}\Bigg[\sum_{j=1,j\neq k}^{M}\sum_{b=1}^{N_b}\alpha_{k,b}^2\bar{\beta}_{j,b}2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t-1}} + N_t\frac{\pi^2}{3}\sum\nolimits_{b=1,b\neq b_k}^{N_b}\alpha_{k,b}^2(1-\bar{\beta}_{k,b})2^{-2B_{k,b}^{\mathrm{PA}}}\Bigg]. \tag{A.19}$$

Such an asymptotical analysis yields an explicit expression for further optimization and analysis, but also removes the impact of user scheduling. In practical systems with limited transmit antennas, the quantized global CDIs of scheduled users are hardly orthogonal. When ZFBF is applied, signal power loss will be caused by projecting the channel of $\mathrm{MS}_k$ to the null-space spanned by the CDIs of the co-scheduled users. To reflect this impact, we introduce a parameter to reflect the average orthogonality, $\mu_k \triangleq \mathbb{E}\{\cos^2\psi_k\}$. Then we can approximate the average SINR estimation as shown in (A.20) at the top of next page.

With (A.20), the QoS requirement of $\mathbb{E}\{\hat{\gamma}_k\} \geq \gamma_k^{\mathrm{QoS}}$ can be derived as the constraint in (8b) after some regular manipulations.

### B. Obtaining Constraint (8c) and (8d)

Because the number of bits should be non-negative and integer-valued, we need exhaustive searching to find the optimal solution, which however does not lead to a closed-form solution. Therefore, we first ignore the integer constraint and only consider the non-negative constraint as shown in (8c) and (8d), then adjust the obtained non-integer bit number to its nearest integer. Together with the constraint (8b), the optimization problem can be formulated as (8).

### C. Finding the Closed-form Solution Shown in (9)

The objective function in (8a) and the constraints in (8c) and (8d) are linear functions of $B_{k,b}^{\mathrm{CDI}}$ and $B_{k,b}^{\mathrm{PA}}$. It is easy to show that the constraint in (8b) is convex with respect to $B_{k,b}^{\mathrm{CDI}}$ and $B_{k,b}^{\mathrm{PA}}$. Therefore, this problem is convex and can be solved from the Karush-Kuhun-Tucker (KKT) conditions, whose solution can be derived as in (9).

## REFERENCES


[1] M. Karakayali, G. Foschini, and R. Valenzuela, "Network coordination for spectrally efficient communications in cellular systems," *IEEE Wireless Commun. Mag.*, vol. 13, no. 4, pp. 56–61, Aug. 2006.

[2] D. Gesbert, S. Hanly, H. Huang, S. Shamai, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: a new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 99, pp. 1380–1408, Dec. 2010.

[3] R. Bhagavatula and R. W. Heath Jr., "Adaptive limited feedback for sum-rate maximizing beamforming in cooperative multicell systems," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 800–811, Feb. 2011.

[4] N. Lee and W. Shin, "Adaptive feedback scheme on K-Cell MISO interfering broadcast channel with limited feedback," *IEEE Trans. Wireless Commun.*, vol. 10, no. 2, pp. 401–406, Feb. 2011.

[5] N. Jindal, "MIMO broadcast channels with finite rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5059, Nov. 2006.

[6] D. Love, R. W. Heath Jr., V. Lau, D. Gesbert, B. 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, Oct. 2008.

[7] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, June 2010.

[8] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1993.

[9] Y. Cheng, V. Lau, and Y. Long, "A scalable limited feedback design for network MIMO using per-cell codebook," *IEEE Trans. Wireless Commun.*, vol. 9, no. 10, pp. 3093–3099, Oct. 2010.

[10] D. Su, X. Hou, and C. Yang, "Quantization based on per-cell codebook in cooperative multi-cell systems," in *Proc. 2011 IEEE Wireless Commun. Netw. Conf.*

[11] J. Zhang and J. Andrews, "Adaptive spatial intercell interference cancellation in multicell wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1–14, Dec. 2010.

[12] R. Zakhour and D. Gesbert, "Optimized data sharing in multicell MIMO with finite backhaul capacity," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6102–6111, Dec. 2011.

[13] D. Love, R. W. Heath Jr., and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.

[14] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1478–1491, Sept. 2007.

[15] S. Han, C. Yang, M. Bengtsson, and A. Perez-Neira, "Channel norm based user scheduling in coordinated multi-point systems," in *Proc. 2009 IEEE Glob. Telecom. Conf.*

[16] A. Wiesel, Y. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4409–4418, Sept. 2008.

[17] J. Zhang, R. Chen, J. Andrews, A. Ghosh, and R. W. Heath Jr., "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910–1921, Apr. 2009.

[18] H. Zhang, N. Mehta, A. Molisch, J. Zhang, and H. Dai, "Asynchronous interference mitigation in cooperative base station systems," *IEEE Trans. Wireless Commun.*, vol. 7, no. 1, pp. 155–165, Jan. 2008.

[19] F. Yuan and C. Yang, "Bit allocation between per-cell codebook and phase ambiguity quantization for limited feedback coordinated multi-point transmission systems," *IEEE Trans. Commun.*, vol. 60, no. 9, pp. 2546–2559, Sept. 2012.

[20] B. Khoshnevis and W. Yu, "Bit allocation laws for multi-antenna channel quantization: multi-user case," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2270–2283, May 2011.

[21] H. Huang, M. Trivellato, A. Hottinen, M. Shafi, P. Smith, and R. Valenzuela, "Increasing downlink cellular throughput with limited network MIMO coordination," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 2983–2989, June 2009.

[22] Q. Zhang and C. Yang, "Semi-dynamic mode selection in base station cooperative transmission system," in *Proc. 2011 IEEE Veh. Technol. Conf. – Fall.*

$$\mathbb{E}\{\hat{\gamma}_k\} \approx \frac{\hat{\mathrm{S}}_{k,\lim}^{\mathrm{LB}}}{\hat{\mathrm{I}}_{k,\lim}^{\mathrm{UB}} + \sigma_k^2} = \frac{\mu_k \sum\limits_{b=1}^{N_b} \alpha_{k,b}^2 (1 - 2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t - 1}}) - \mu_k \frac{\pi^2}{3} \sum\limits_{b=1, b\neq b_k}^{N_b} \alpha_{k,b}^2 \left(1 - \bar{\beta}_{k,b}\right) 2^{-2B_{k,b}^{\mathrm{PA}}}}{\sum\limits_{j=1, j\neq k}^{M} \sum\limits_{b=1}^{N_b} \frac{1}{N_t} \alpha_{k,b}^2 \bar{\beta}_{j,b} 2^{-\frac{B_{k,b}^{\mathrm{CDI}}}{N_t - 1}} + \frac{\pi^2}{3} \sum\limits_{b=1, b\neq b_k}^{N_b} \alpha_{k,b}^2 (1 - \bar{\beta}_{k,b}) 2^{-2B_{k,b}^{\mathrm{PA}}} + \frac{M\sigma_k^2}{N_b N_t P_0}}. \quad \text{(A.20)}$$


[23] F. Boccardi and H. Huang, "Optimum power allocation for the MIMO-BC zero-forcing precoder with per-antenna power constraints," in *Proc. 2006 Conf. Inf. Sciences Syst.*

[24] 3GPP Long Term Evolution (LTE), "Further advancements for E-UTRA physical layer aspects," TSG RAN TR 36.814 v9.0.0, Mar. 2010.

[25] C. Murthy and B. Rao, "Quantization methods for equal gain transmission with finite rate feedback," *IEEE Trans. Signal Process.*, vol. 55, no. 1, pp. 233–245, Jan. 2007.

[26] O. Somekh, O. Simeone, Y. Bar-Ness, A. M. Haimovich, and S. Shamai, "Cooperative multicell zero-forcing beamforming in cellular downlink channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3206–3219, July 2009.


**Xueying Hou** received the B.S. degree from Beihang University, China, in 2007. She is currently working toward the Ph.D. degree in signal and information processing with the School of Electronics and Information Engineering, Beihang University. From September 2009 to June 2010, she was a Visiting Student with the School of Electrical Engineering, Royal Institute of Technology (KTH), Stockholm, Sweden. Her research interests include cooperative communication, channel estimation and limited feedback techniques in wireless systems.

**Chenyang Yang** (SM'08) received the M.S.E and Ph.D. degrees in electrical engineering from Beihang University (formerly Beijing University of Aeronautics and Astronautics), Beijing, China, in 1989 and 1997, respectively.

She is currently a Full Professor with the School of Electronics and Information Engineering, Beihang University. She has published various papers and filed many patents in the fields of signal processing and wireless communications. Her recent research interests include signal processing in network MIMO, cooperative communication, energy efficient transmission and interference management.

Prof. Yang was the Chair of the IEEE Communications Society Beijing chapter from 2008 to 2012. She has served as Technical Program Committee Member for many IEEE conferences, such as the IEEE International Conference on Communications and the IEEE Global Telecommunications Conference. She currently serves as an Associate editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Associate Editor-in-Chief of the *Chinese Journal of Communications*, and an Associate editor-in-chief of the *Chinese Journal of Signal Processing*. She was nominated as an Outstanding Young Professor of Beijing in 1995 and was supported by the First Teaching and Research Award Program for Outstanding Young Teachers of Higher Education Institutions by Ministry of Education (P.R.C. "TRAPOYT") during 1999-2004.