A Limited-Feedback Scheduling and Beamforming Scheme for Multi-User Multi-Antenna Systems

Behrouz Khoshnevis and Wei Yu
Department of Electrical and Computer Engineering
University of Toronto, Toronto, Ontario, Canada
Email: {bkhoshnevis, weiyu}@comm.utoronto.ca

***Abstract*—This paper proposes an efficient two-stage limited-feedback beamforming and scheduling scheme for multiple-antenna cellular communication systems. The system model includes a base-station with M antennas and a large pool of users with a total feedback rate of B bits per fading block. The feedback process is divided into two stages. In the first stage, the users measure their channel gains from each antenna and feedback the index of the antenna with the highest channel gain along with the gain itself. Based on this information, the base-station schedules M users with the highest channel gains from its M antennas and polls those users for explicit quantization of their vector channels in the second stage. Based on these quantized channels, the base-station then forms zero-forcing beamforming vectors for downlink transmission. This paper presents an approximate analysis for the proposed scheme which is used to optimize the bit allocation between the two feedback stages. It is shown that for a total number of feedback bits B, the number of feedback bits assigned to the second stage, B_2, should scale as $M(M-1)\log(\mathtt{SNR} \times B)$. In particular, the fraction B_2/B behaves as $\log B/B$ in the asymptotic regime where $B \to \infty$. Further, the approximate downlink sum rate is shown to scale as $M \log \mathtt{SNR} + M \log\log B$, suggesting that both multiuser multiplexing and multiuser diversity gains are realized. As the numerical results verify, the proposed feedback scheme, in spite of its low complexity, performs very close to the more complicated beamforming and scheduling schemes in the literature and in fact outperforms such schemes in the high-SNR regime.**

I. Introduction

The advantage of multi-user multi-antenna systems lies in their promise in achieving both spatial multiplexing and multi-user diversity gains. The realization of these gains, however, depends critically on the availability of users' channel state information (CSI) at the base-station. Acquiring CSI is a challenging issue especially in *limited-feedback systems*, where users need to explicitly quantize and feedback their channel information through a rate-limited feedback channel. Due to the scarcity of the feedback capacity in practical systems, the design of beamforming and scheduling algorithms that can efficiently utilize the feedback bandwidth introduces an interesting challenge, and has attracted a great deal of research recently [1]–[12].

Most of the multi-user scheduling and beamforming algorithms in the literature fall into one of the following two categories. The first line of work, as in [1]–[3], assumes fixed orthogonal beamforming codebooks. The users feedback the index of the beam with the highest signal-to-interference-plus-noise ratio (SINR) along with the corresponding SINR value. The base-station then selects the user with the highest SINR on each beam and eventually uses the same orthogonal beams for downlink transmission. We refer to this approach as the *orthogonal beamforming* (OBF) approach in this paper.

In the second approach, as in [4]–[6], the users explicitly quantize and feedback the channel direction information (CDI) along with certain channel quality indicators (CQI). The base-station then uses this information for scheduling and beamforming. One of the well-known and practically feasible scheduling-beamforming algorithms is the greedy user selection with zero-forcing beamforming [4]. This combined scheduling-beamforming approach is referred to as the *zero-forcing beamforming* (ZFBF) approach.

Each of these two schemes have their merits and disadvantages. The main advantage of the OBF approach lies in the simplicity of its scheduling algorithm. The OBF scheme, however, suffers from the low accuracy of the quantized channel information. In order to improve the accuracy of the quantized information, a variation of the OBF approach is presented by [7], where a collection of orthogonal codebooks is used instead of a single codebook. The authors of [6] numerically compare the performance of ZFBF with the performance of OBF proposed by [7] in terms of the downlink sum rate under a total feedback rate constraint. The comparison reveals that ZFBF outperforms OBF for almost any feedback rate constraint. As OBF is easier to implement than ZFBF, there appears to be a tradeoff between the superior performance of ZFBF and the lower computational complexity of OBF.

This paper proposes a two-stage feedback mechanism that achieves a performance comparable to the ZFBF scheme with a scheduling complexity comparable to the OBF scheme. The main idea is to decompose the feedback process into two stages that are separately used for scheduling and beamforming. The first stage is similar to the OBF scheme, where the base-station schedules users based on their SINR feedback values. In the second stage, the scheduled users are asked to explicitly quantize and feedback their channel directions. The base-station then uses the quantized directions to form zero-forcing beamforming vectors that are eventually used for downlink transmission. The proposed scheme is shown to have similar performance as ZFBF scheme with far less computational complexity. Such an advantage makes the proposed feedback mechanism a powerful candidate for practical

implementations.

We should mention that the idea of two-stage feedback is originally proposed by the authors of [8]. The algorithm in [8] however uses greedy user scheduling and has the same computational complexity as ZFBF. Our approach, on the other hand, offers an OBF-like scheduling complexity with a performance at least as good as ZFBF.

Finally, we comment that the beamforming-scheduling algorithms discussed here are deterministic in nature, i.e., there is no probabilistic contention between users in accessing the feedback channel. For a discussion on contention-based scheduling algorithms, the reader is referred to [9] and [10].

II. System Model

Consider a base-station with M antennas and a pool of users indexed by k. User channels $\mathbf{h}_k$ are i.i.d. with $\mathcal{CN}(0,1)$ entries. The users have perfect knowledge of their own channels and provide CSI back to the base-station through a feedback channel with a total number of B feedback bits per fading block.

For scheduling, we adopt a similar approach as in OBF scheme and use the columns of a $M\times M$ identity matrix as the scheduling orthogonal beams[1]. User k's channel gain along m'th beam is therefore simply the m'th entry of the channel vector $\mathbf{h}_k$, which is denoted by $\mathbf{h}_{k,m}$. To allow the users to measure their channel gains from each antenna, the base-station transmits pilot signals prior to the feedback process.

The feedback process, as shown in Fig. 1, is divided into two stages using B_1 and B_2 bits respectively. In the first stage, users feedback the index of the antenna with the highest channel gain along with the gain itself[2]. The base-station then chooses the user with the highest channel gain from each antenna:

$$\pi(m) = \arg\max_k |\mathbf{h}_{k,m}|, \tag{1}$$

where $1\leq m\leq M$ and $\pi(m)$ is the index of the user scheduled for the m'th antenna. Given that the number of feedback bits for scheduling stage is B_1, the number of users allowed to participate is[3]

$$N = \left\lfloor \frac{B_1}{\log M} \right\rfloor, \tag{2}$$

which are randomly chosen from the pool of users.

In the second stage, the M scheduled users quantize and feedback their channel directions $\hat{\mathbf{h}}_k = \mathbf{h}_k/\|\mathbf{h}_k\|$. With a total of B_2 bits in the second stage, each user uses

$$b = \left\lfloor \frac{B_2}{M} \right\rfloor \tag{3}$$

bits for quantization. The base-station then uses the quantized directions to form zero-forcing beamforming vectors $\mathbf{v}_m$, $1\leq m\leq M$, which are used for downlink transmission.

[1]With sufficient user mobility, fixing the scheduling beams would not degrade the scheduling fairness.

[2]Note that the gain information is assumed to be unquantized for simplicity. Similarly, in simulating the ZFBF scheme in Section IV, we assume that the channel quality indicators (CQI) are unquantized.

[3]All log functions in this paper are base-2.

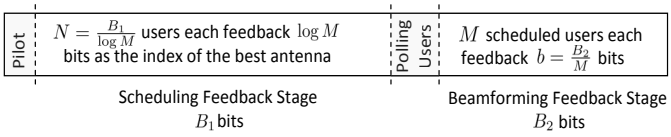

Fig. 1. Two-stage feedback process.

Assuming that the transmission power, denoted by SNR, is equally divided among the scheduled users, the expected downlink sum rate is given by

$$R = M\mathbb{E}\left[\log\left(1+\frac{\rho\left\|\mathbf{h}_{\pi(m)}\right\|^2\left|\hat{\mathbf{h}}_{\pi(m)}^{\dagger}\mathbf{v}_m\right|^2}{1+\sum\limits_{n\neq m}\rho\left\|\mathbf{h}_{\pi(m)}\right\|^2\left|\hat{\mathbf{h}}_{\pi(m)}^{\dagger}\mathbf{v}_n\right|^2}\right)\right], \tag{4}$$

where $\rho = \frac{\mathrm{SNR}}{M}$.

Our objective is to maximize the sum rate subject to the total feedback rate constraint:

$$\max_{B_1,B_2} \quad R(B_1,B_2) \tag{5}$$

$$\text{s.t.} \quad B_1+B_2 = B. \tag{6}$$

In order to understand the dependence of the sum rate on B_1 and B_2, consider the expression in (4). As B_1 increases and more users participate in the scheduling stage, we have a better chance of finding users with higher channel gains $\left\|\mathbf{h}_{\pi(m)}\right\|$. On the other hand, as B_2 increases, the scheduled users can provide more accurate quantization of their channel directions $\hat{\mathbf{h}}_{\pi(m)}$ and the zero-forcing beamforming vectors $\mathbf{v}_m$ would be more efficient in removing the multi-user interference in the denominator of the rate expression. The next section uses an approximate analysis of the sum rate to optimize the bit allocation between B_1 and B_2 and studies the system performance with such bit allocation.

III. Approximate System Analysis

First, we note that $\left\|\mathbf{h}_{\pi(m)}\right\|^2 \leq M\left|\mathbf{h}_{\pi(m),m}\right|^2$, since user $\pi(m)$'s channel gain from the m'th antenna is stronger than its channel gain from other antennas. Combining this with (4), we achieve the following upper bound for the sum rate:

$$\begin{aligned} R &\leq M\mathbb{E}\left[\log\left(1+\frac{M\rho\left|\mathbf{h}_{\pi(m),m}\right|^2\left|\hat{\mathbf{h}}_{\pi(m)}^{\dagger}\mathbf{v}_m\right|^2}{1+\sum\limits_{n\neq m}M\rho\left|\mathbf{h}_{\pi(m),m}\right|^2\left|\hat{\mathbf{h}}_{\pi(m)}^{\dagger}\mathbf{v}_n\right|^2}\right)\right] \\ &\leq M\log\left(\mathbb{E}\left[1+\frac{M\rho\left|\mathbf{h}_{\pi(m),m}\right|^2\left|\hat{\mathbf{h}}_{\pi(m)}^{\dagger}\mathbf{v}_m\right|^2}{1+\sum\limits_{n\neq m}M\rho\left|\mathbf{h}_{\pi(m),m}\right|^2\left|\hat{\mathbf{h}}_{\pi(m)}^{\dagger}\mathbf{v}_n\right|^2}\right]\right) \end{aligned} \tag{7}$$

where we have used Jensen's inequality.

Unfortunately, finding a closed-form expression for the upper bound in (7) appears to be difficult (if not impossible). We therefore consider an approximation of the sum rate by replacing each of the random terms with its expected value.

Similar approaches are used in the earlier literature, e.g. [6]. A justification of this approximate method is presented in the appendix.

The bit allocations that result from such an approximation clearly are suboptimal. Nevertheless, as the numerical results in the next section verify, the proposed analysis provides a reasonably accurate approximation of the system performance. Particularly, the asymptotic bit allocations that result from this analysis are highly accurate when compared with the optimal bit allocations achieved through simulation.

The following describes our approximation of the upper bound expression in (7). First, we note that by using orthogonal beams for scheduling, the scheduled users are guaranteed to have small spatial correlation, therefore the zero-forcing beamforming vectors are expected to be nearly aligned with users' channels. With this justification, we use the following approximation:

$$\mathbb{E}\left[\left|\hat{\mathbf{h}}_{\pi(m)}^{\dagger}\mathbf{v}_m\right|^2\right] \approx 1. \tag{8}$$

Next, for the interference term, we use the following approximation from [5]:

$$\mathbb{E}\left[\left|\hat{\mathbf{h}}_{\pi(m)}^{\dagger}\mathbf{v}_n\right|^2\right] \approx \frac{1}{M-1}2^{-b/(M-1)} \approx \frac{1}{M-1}2^{-B_2/M(M-1)}, \tag{9}$$

where the second approximation ignores the floor function (3). The expression in (9) is exact if the channel directions $\hat{\mathbf{h}}_{\pi(m)}$ are independent and uniformly distributed on the complex unit hypersphere. However, the scheduled users' channel directions are neither independent nor uniformly distributed. The expression in (9) is therefore only an approximation.

Finally, according to the scheduling rule in (1), we have

$$\left|\mathbf{h}_{\pi(m),m}\right| = \max_{1\le k\le N}\left|\mathbf{h}_{k,m}\right|, \tag{10}$$

where N is the number of users participating in the scheduling stage. Therefore, $\left|\mathbf{h}_{\pi(m),m}\right|^2$ is the maximum of N independent $\chi^2(2)$ random variables and its expectation, according to [1], behaves as $\ln(N)$:

$$\mathbb{E}\left[\left|\mathbf{h}_{\pi(m),m}\right|^2\right] \approx \ln(N) \approx \ln(B_1/\log M). \tag{11}$$

The second approximation in (11) ignores the floor function in (2).

By substituting the average values in (8), (9), and (11) in the rate function in (7), we achieve the following approximation for the sum rate:

$$\tilde{R} = M\log\left(1+\frac{M\rho\ln(B_1/\log M)}{1+M\rho\ln(B_1/\log M)2^{-\frac{B_2}{M(M-1)}}}\right). \tag{12}$$

By optimizing $\tilde{R}$ with respect to B_1 and B_2 subject to the constraint in (6), we arrive at the following equations for B_1 and B_2:

$$\begin{cases} \frac{\rho\ln 2}{M-1}B_1(\ln B_1-\mu)^2 = 2^{\frac{B_2}{M(M-1)}} \\ B_1+B_2 = B \end{cases} \tag{13}$$

where $\mu = \ln\log M$.

We can further simplify the problem by assuming an asymptotic regime where $B\to\infty$. By taking the logarithm of both sides of (13), we have

$$\log B_1 + \log\rho \doteq \log B_1 + 2\log(\ln B_1-\mu) + \eta = \frac{B_2}{M(M-1)} \tag{14}$$

where $\eta = \log\frac{\rho\ln 2}{M-1}$ and the notation $f(B)\doteq g(B)$ means $\lim_{B\to\infty} f(B)/g(B) = 1$. Combining (14) with the constraint $B_1+B_2=B$ we arrive at $B_1\doteq B$. Substituting this back in (14), we get the following asymptotic bit allocation:

$$B_1 \doteq B \tag{15}$$

$$B_2 \doteq M(M-1)\log(\rho B). \tag{16}$$

The asymptotic results in (15) and (16) show that as the total feedback rate increases, higher percentage of bits are used for the scheduling stage. In particular, the percentage of bits used for the beamforming stage behaves as $\log B/B$ as $B\to\infty$.

Finally, by substituting the asymptotic bit allocations in the approximate rate expression in (12), one can easily show that

$$\tilde{R} \doteq M\log(1+\rho\ln B) \doteq M\log\rho + M\log\log B, \tag{17}$$

which suggests that both multiplexing gain and multi-user diversity gain are realized.

IV. Numerical Results

This section compares the numerical results achieved through simulation by those suggested by the proposed approximate analysis. Users' channel vectors in simulations are assumed to be independent with i.i.d. complex Gaussian $\mathcal{CN}(0,1)$ entries as stated in Section II.

We start with the bit allocation results. Fig. 2 shows the optimal percentage of bits allocated to the beamforming feedback stage in the proposed two-stage feedback scheme for a system with $M=4$ antennas and $\mathtt{SNR}=15$dB. As the asymptotic bit allocation result in (16) suggests, this percentage scales as $\log B/B$ as B increases.

Fig. 3 shows the same percentage as a function of $\mathtt{SNR}$, when the total number of feedback bits is fixed at $B=300$ bits. As (16) suggests, the number of feedback bits allocated to CSI quantization in the second stage scales linearly with $\mathtt{SNR}$ in dB scale. This coincides with the result in [5], which states that the number of feedback bits per user should scale as $(M-1)\log\mathtt{SNR}$ in order to preserve the multiplexing gain in a network of M users. Of course, this scaling will saturate at some point, since the total number of feedback bits is fixed.

Next, we compare the performance of the proposed two-stage feedback scheme with the performance of ZFBF scheme. In ZFBF scheme, K users each feedback B/K bits. The base-station then selects M users out of these K users using greedy user scheduling. The performance of this scheme is optimized over the number of users K that participate in feedback much like in [6].

Fig. 4 and Fig. 5, show the performances of the ZFBF and two-stage feedback schemes achieved through simulation

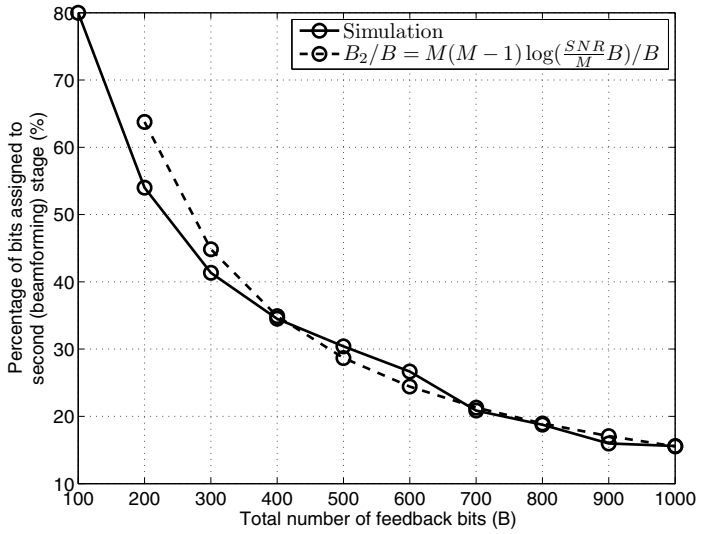

Fig. 2. Optimal bit allocation to the beamforming feedback stage for a system with $M = 4$ antennas and $\mathtt{SNR} = 15$dB.

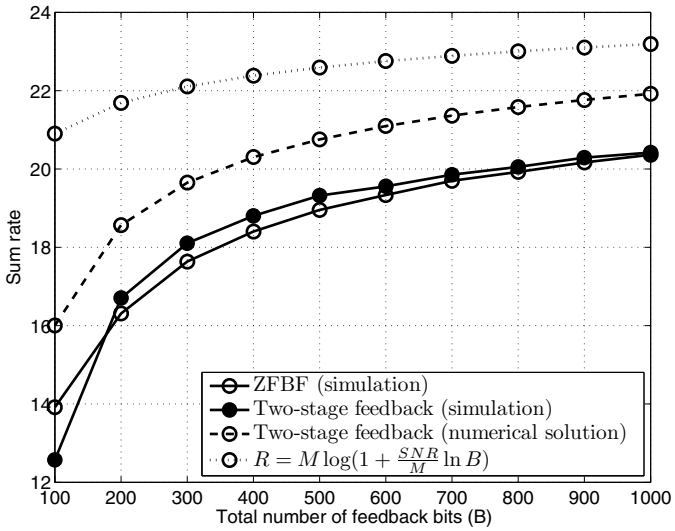

Fig. 4. Sum rate as a function of the total feedback rate for a system with $M = 4$ antennas and $\mathtt{SNR} = 15$dB.

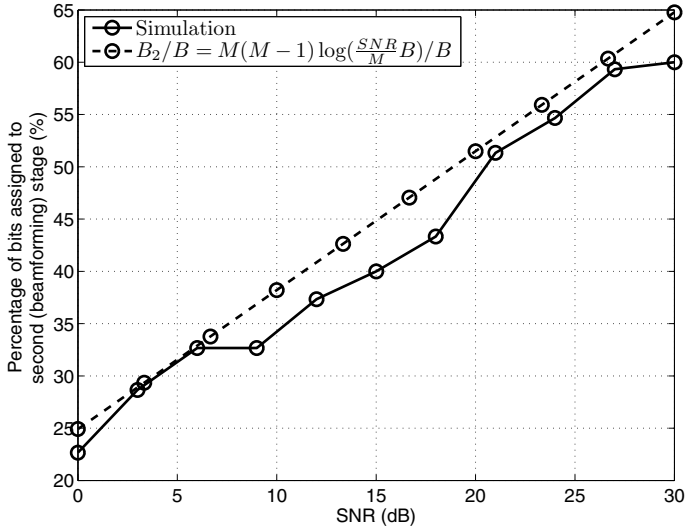

Fig. 3. Optimal bit allocation to the beamforming feedback stage for a system with $M = 4$ antennas and $B = 300$ bits.

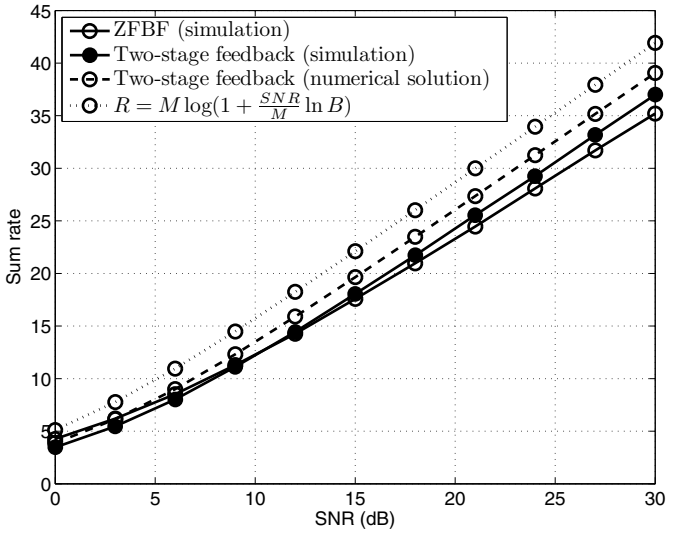

Fig. 5. Sum rate as a function of SNR for a system with $M = 4$ antennas and $B = 300$ bits.

for fixed values of $\mathtt{SNR}$ and B respectively. The figures also plot the performance of the two-stage feedback scheme when one uses the numerical solution of equations in (13) for bit allocation between the scheduling and beamforming stages. The results show that the ZFBF scheme and the two-stage feedback scheme have similar performances, with the two-stage scheme slightly outperforming ZFBF in high-SNR regime. The two figures also include the asymptotic sum-rate upper bound in (17). This upper bound appears to have an offset of almost 15% in predicting the actual sum rate; however, it accurately projects the logarithmic and double logarithmic scaling of the actual sum rate with $\mathtt{SNR}$ and B respectively.

In order to investigate the computational complexity of the proposed scheme, Fig. 6 presents typical CPU processing times required for scheduling computations. As the figure shows, the two-stage algorithm is almost 10-20 times faster in comparison to the ZFBF greedy user selection approach. Considering the performance similarity between the proposed scheme and the ZFBF scheme, the far less computational complexity of the two-stage scheme makes it a more favorable candidate for practical system implementations.

As a final note, we mention that the two-stage feedback scheme imposes an additional delay on the feedback process due to the two phases involved. The proposed scheme is therefore most suited for systems with sufficiently large channel coherence time.

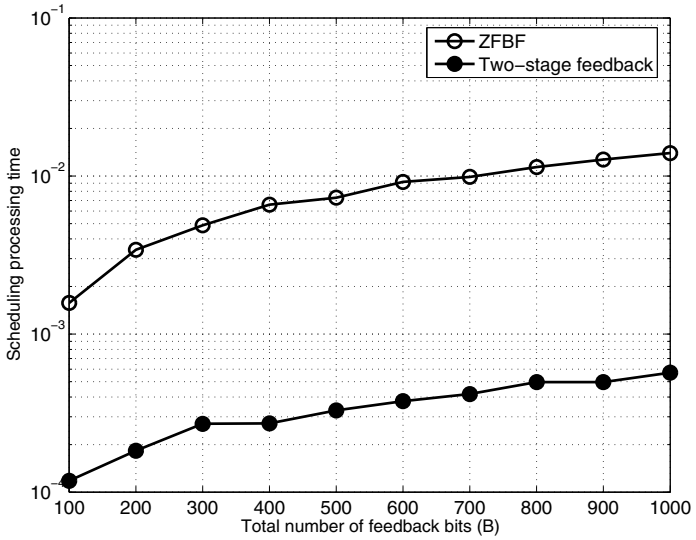

Fig. 6. CPU processing time required for scheduling computations for a system with $M = 4$ antennas and $\mathtt{SNR} = 15$dB (CPU: Intel Xeon 2.33GHZ).

V. Conclusions

This paper proposes a two-stage feedback mechanism for limited-feedback cellular systems. In the first stage, the base-station schedules users based on their channel gains from the transmission antennas. The scheduled users are then polled to feedback their CSI in the second stage. The optimal bit allocation between the two stages is derived using an approximate analysis. The proposed two-stage feedback scheme has a far less computational complexity with a performance at least as good as other available scheduling-beamforming schemes in the literature. These advantages make the proposed scheme a powerful candidate for practical system implementations.

Appendix

This section provides a justification for the approximation process in Section III, where we replace each term in the upper bound in (7) with its expected value. To this end, define

$$X_{m,n} = \left|\mathbf{h}_{\pi(m),m}\right|^2 \left|\hat{\mathbf{h}}^{\dagger}_{\pi(m)}\mathbf{v}_n\right|^2, \tag{18}$$

where $1 \leq m, n \leq M$. The upper bound in (7) can therefore be written as follows:

$$R < M \log\left(\mathbb{E}\left[1+\frac{\mathtt{SNR}X_{m,m}}{1+I_m}\right]\right), \tag{19}$$

where I_m is the interference power:

$$I_m = \mathtt{SNR}\sum_{n\neq m} X_{m,n}. \tag{20}$$

In the asymptotic regime, where $B \to \infty$, any efficient feedback mechanism should force the interference power to zero; otherwise, the rate expression would saturate in high-SNR regime. For example, for the feedback scheme proposed in this paper, if we consider the interference term in the denominator of (12) and apply the bit allocation rules in (15) and (16), we see that the interference power diminishes as $\ln B/B$ as $B \to \infty$.

We therefore expect $X_{m,n}\to 0$, for $n\neq m$, as $B\to\infty$. Furthermore, our numerical results suggest that, for the scheduling-beamforming scheme in this paper, the ratio

$$\frac{\sigma^2_{X_{m,n}}}{\mathbb{E}[X^2_{m,n}]} = \frac{\mathbb{E}\left[\left|X_{m,n}-\mathbb{E}[X_{m,n}]\right|^2\right]}{\mathbb{E}[X^2_{m,n}]} \tag{21}$$

also diminishes as $B \to \infty$. This result, although difficult to prove due to the complexity of $X_{m,n}$'s probability distribution function, suggests that one can safely ignore the difference term $(X_{m,n} - \mathbb{E}[X_{m,n}])$ in comparison with $\mathbb{E}[X_{m,n}]$ and therefore safely use the approximation $X_{m,n} \approx \mathbb{E}[X_{m,n}]$.

Using this justification, we can approximate the SINR terms in (19) as follows:

$$\mathbb{E}\left[1+\frac{\mathtt{SNR}X_{m,m}}{1+\mathtt{SNR}\sum\limits_{n\neq m} X_{m,n}}\right] \approx \mathbb{E}\left[1+\frac{\mathtt{SNR}X_{m,m}}{1+\mathtt{SNR}\sum\limits_{n\neq m} \mathbb{E}[X_{m,n}]}\right] = 1+\frac{\mathtt{SNR}\mathbb{E}[X_{m,m}]}{1+\mathtt{SNR}\sum\limits_{n\neq m}\mathbb{E}[X_{m,n}]}$$

which justifies replacing each random variable in the upper bound in (7) with its expected value.

References

[1] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side infonnation," *IEEE Trans. Inform. Theory*, vol. 51, no. 2, pp. 506-522, Feb. 2005.

[2] K. Huang, R. Heath Jr., and J. Andrews, "Space division multiple access with a sum feedback rate constraint," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3879-3891, Jul. 2007.

[3] J. S. Kim, H. Kim, and K. B. Lee, "Limited feedback signaling for MIMO broadcast channels," in *Proc. IEEE Int. Workshop on Signal Proc. Adv. Wireless Commun.*, June 2005, pp. 855-859.

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

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

[6] N. Ravindran and N. Jindal, "Multi-user diversity vs. accurate channel state information in MIMO downlink channels," submitted for publication. [Online]. Available: http://arxiv.org/abs/0907.1099

[7] K. Huang, J. Andrews, and R. Heath, Jr., "Performance of orthogonal beamforming for SDMA with limited feedback," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 152-164, Jan. 2009.

[8] R. Zakhour and D. Gesbert, "A two-stage approach to feedback design in multi-user MIMO channels with limited channel state information," in *Proc. IEEE Int. Symp. Personal, Indoor, and Mobile Radio Commun.*, Sept. 2007.

[9] R. Agarwal, C. Hwang, and J. Cioffi, "Opportunistic feedback protocol for achieving sum-capacity of the MIMO broadcast channel," in *Proc. IEEE Veh. Technol. Conf.*, Sept.-Oct. 2007, pp. 606-610.

[10] T. Tang, R. Heath, Jr., S. Cho, and S. Yun, "Opportunistic feedback for multiuser MIMO systems with linear receivers," *IEEE Trans. Commun.*, vol. 55, no. 5, pp. 1020-1032, May 2007.

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

[12] B. Khoshnevis and W. Yu, "Bit allocation laws for multi-antenna channel quantization: multi-user case," to appear in *IEEE Trans. Signal Process.*, 2011.