MIMO Broadcast Scheduling with Quantized Channel State Information

Charles Swannack, Gregory W. Wornell
Dept. EECS, MIT
Cambridge MA 02139
Email: {swannack,gww}@mit.edu

Elif Uysal-Biyikoglu
Dept. ECE, Ohio State University
Columbus, OH 43210
Email: elif@ece.osu.edu

***Abstract*— We develop and analyze a simple, low-complexity system architecture for scheduling over a Gaussian multiple-input multiple-output (MIMO) broadcast channel with infinite message backlogs. In the system of interest, there is a transmitter with m antennas, and n receiving users, where $n \gg m$. We show that the proposed architecture is strongly asymptotically optimal with respect to average throughput. We further characterize the feedback requirements of the architecture, and highlight various tradeoffs available to the system designer.**

I. Introduction and Background

There is growing interest in the development of efficient wireless broadcast systems for distributing independent data streams to different users over some geographical area. It is now widely appreciated that the use of a multiple-element antenna array at the transmitter can, in principle, greatly increase the capacity of such systems; see, e.g., [1]. When the number of users is no larger than the array size, the system design issues are rather well-understood. Moreover, when it is desirable for complexity or other reasons to restrict one's attention to case of linear multiplexing, the literature characterizing the associated performance tradeoffs is particularly extensive; see, e.g., [2]–[4].

By contrast, comparatively little is known about how to design efficient systems when the number of users becomes large relative to the array size, and in particular the nature of the fundamental tradeoffs between throughput, complexity, and feedback in such settings. Ultimately, the underlying scheduling problem is rather different and in many ways richer than that of more traditional networks.

There is a growing literature on the problem of MIMO scheduling — see, e.g. [5] and references there in. Within this domain, much of the recent work has focused on examining the throughput scaling behavior in the large user pool regime under various system complexity constraints and with the assumption of perfect channel channel state information at the transmitter [6]–[9]. In this paper, we develop a simple feedback-based scheduling architecture and establish that it achieves a strong form of asymptotic throughput optimality.

II. Channel and System Model

The system of interest consists of an m-element transmitter antenna array and a pool of n destinations (users). The transmitter has n collections of messages, each such collection destined for one of the n users. The collections are infinite in size, corresponding to an infinite backlog.

Our discrete-time channel model is a narrowband block fading one. Specifically, in any particular block, the signal $y_j(k)$ received by user j at time k in response to a signal $\mathbf{x}(k)$ transmitted from the array is of the form

$$y_j(k) = \mathbf{h}_j^\dagger \mathbf{x}(k) + z_j(k) \tag{1}$$

where $z_j(k)$ is independent identically distributed (i.i.d.) $\mathcal{CN}(0,1)$ noise, and where the (normalized) channel gain vectors $\mathbf{h}_j$ have i.i.d. $\mathcal{CN}(0,1/2m)$ elements. The noises and channel gains are independent from receiver to receiver, and from block to block.

Any message scheduled for delivery is transmitted within one block, and the blocks are long so that messages can be reliably received. Thus each block corresponds to a new signaling (and hence scheduling) interval. Within each signaling interval, the transmitter sends from its array a group of messages, one for each of a subset of the user pool. The transmitter is subject to an average total power constraint of P, i.e., $E\left[\|\mathbf{x}\|^2\right] \leq P$ within each signaling interval.

In our model, channel gains in each signaling interval are known perfectly (i.e., measured to arbitrary accuracy) at the respective receivers at the beginning of each such interval. Moreover, a feedback link exists by which individual users can inform the transmitter of their channel gains (or more generally quantized versions thereof), also at the beginning of each associated signaling interval. The users do not know each other's channel gains, nor are they able to more generally share information between each other.

Finally, the performance criterion of interest in this work is average throughput (i.e., expected sum-rate), and our focus is on the large n regime (with m fixed).

III. System and Protocol Architecture

The architecture of interest is as illustrated in Fig. 1. The protocol is identical in each signaling interval, so we restrict our attention to a single arbitrary one.

In such an interval, a subset $\mathcal{R}$ of users from the full population $\mathcal{U}$ send a quantized representation of their respective channel gain vectors to the transmitter over the feedback link. The associated quantization codebook $\mathcal{C}$ is fixed and the same

This work was supported in part by NSF under Grant No. CNS-0434974, Mitre Corporation, and by HP through the MIT/HP Alliance.

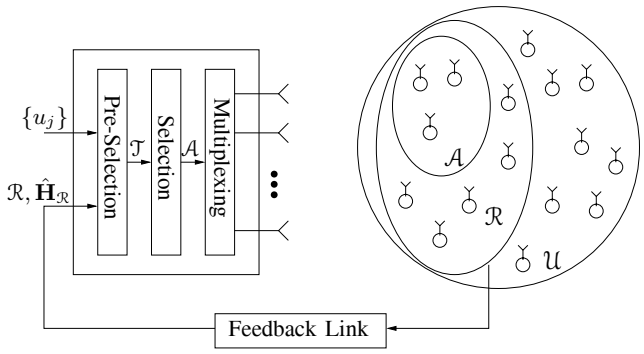

Fig. 1. MIMO system architecture. In each scheduling interval a subset $\mathcal{R}$ of the full user pool $\mathcal{U}$ of size n reports quantizations $\hat{\mathbf{H}}_{\mathcal{R}}$ of its channel gains to the transmitter via the feedback link using a decentralized (individual) criterion. From the set $\mathcal{R}$, the transmitter first forms a collection $\mathcal{T}$ of candidate user sets of size m using a pairwise criterion; this is the pre-selection phase. Next, a set $\mathcal{A} \in \mathcal{T}$ is chosen at random as the active set, whose messages $\{u_j, j \in \mathcal{A}\}$ are linearly multiplexed across the array for transmission.

for all users. Its structure is such that the codewords $\mathbf{c} \in \mathcal{C}$ all lie on the unit sphere in m (complex) dimensions, and the quantization rule corresponds to

$$\hat{\mathbf{h}}_j = \arg\max_{\mathbf{c}\in\mathcal{C}} \left|\mathbf{c}^\dagger \mathbf{h}_j\right|, \tag{2}$$

where $\hat{\mathbf{h}}_j$ denotes the quantization of $\mathbf{h}_j$. We let r denote the number of bits to which a channel gain is quantized, so the the codebook is of size 2^r. We label the codewords in the codebook $\mathcal{C} = \mathcal{C}_r$ as $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{2^r}$. With this notation a key figure of merit for the codebook is its *coherence*

$$\mu(\mathcal{C}) = \max_{i\neq j} \left|\mathbf{c}_i^\dagger \mathbf{c}_j\right|. \tag{3}$$

In general, $0 \leq \mu \leq 1$, and, for a given r, smaller values of μ correspond to better codes.

The subset $\mathcal{R}$ is determined in a decentralized manner, i.e., based on an individual evaluation of each channel gain vector. Specifically, each user j computes the squared norm $\|\mathbf{h}_j\|^2$ of its channel gain vector, and the correlation $|\mathbf{h}_j^\dagger \hat{\mathbf{h}}_j|$ between the channel gain vector and its quantization $\hat{\mathbf{h}}_j$. If these factors fall within certain prescribed ranges, a user will convey its channel gain to the transmitter. The particular criterion we consider corresponds to

$$\mathcal{R}_{\rho,\sigma} \triangleq \left\{j \in \mathcal{U} : \rho^- \leq \|\mathbf{h}_j^2\| \leq \rho^+ \text{ and } |\tilde{\mathbf{h}}_j^\dagger \hat{\mathbf{h}}_j| \geq \sigma\right\}, \tag{4}$$

where $\tilde{\mathbf{h}}_j = \mathbf{h}_j/\|\mathbf{h}_j\|$, and where ρ^+, ρ^-, and σ are prescribed parameters of the protocol. Furthermore, it suffices to restrict our attention to $\sigma \geq \mu(\mathcal{C})$. In the special case where no quantization is used, the criterion (4) specializes to that corresponding to

$$\mathcal{R}_{\rho} \triangleq \left\{j \in \mathcal{U} : \rho^- \leq \|\mathbf{h}_j\|^2 \leq \rho^+\right\}. \tag{5}$$

At the transmitter, there are three relevant stages of processing. First, from the set $\mathcal{R}$ of reporting users, a collection $\mathcal{T}$ of candidate subsets of size[1] m is formed; this is the pre-selection phase. Next, one of these subsets, denoted $\mathcal{A}$, is selected from $\mathcal{T}$ at random, and corresponds to the active user set for the signaling interval. Finally, one message for each of the active users is selected, and the resulting group of messages is multiplexed across the array for transmission.

The pre-selection phase is based on simple pairwise evaluation of the vectors in $\mathcal{R}$. The particular criterion we consider corresponds to

$$\mathcal{T}_{\epsilon,\rho,\sigma} \triangleq \left\{\mathcal{A} \subset \mathcal{R}_{\rho,\sigma} : |\mathcal{A}| = m \text{ and } |\hat{\mathbf{h}}_i^\dagger \hat{\mathbf{h}}_j| \leq \epsilon,\ \forall\, i \neq j \in \mathcal{A}\right\}, \tag{6}$$

where ϵ is another prescribed parameter of the protocol. In the special case where no quantization is used, the criterion (6) specializes to

$$\mathcal{T}_{\epsilon,\rho} \triangleq \left\{\mathcal{A} \subset \mathcal{R}_{\rho} : |\mathcal{A}| = m \text{ and } |\tilde{\mathbf{h}}_i^\dagger \tilde{\mathbf{h}}_j| \leq \epsilon,\ \forall\, i \neq j \in \mathcal{A}\right\}. \tag{7}$$

In general, the parameters ρ^+, ρ^-, σ, and ϵ — and hence the sets $\mathcal{R}$, $\mathcal{T}$, and $\mathcal{A}$ — will all be functions of n. Our notation will only show this dependency explicitly when necessary.

For the multiplexing phase of the protocol, we restrict our attention to linear multiplexers. Specifically, with $\mathbf{u}$ denoting the vector of m coded symbols $u_j, j \in \mathcal{A}$ for the m active users, the transmitted signal takes the form

$$\mathbf{x} = \sum_{j\in\mathcal{A}} u_j \mathbf{w}_j = \mathbf{W}_{\mathcal{A}} \mathbf{u} \tag{8}$$

where the $\mathbf{W}$ is a matrix whose columns are the unit-norm weight (i.e., beamforming) vectors $\mathbf{w}_j, j \in \mathcal{A}$. We further restrict our attention to uniform power allocation in the multiplexing, i.e., $E\left[|u_j|^2\right] = P/m$ for all $j \in \mathcal{A}$.

Among linear multiplexers, of primary interest will be interference-cancelling (IC) — i.e., zero-forcing — multiplexers. The weight matrix in this case takes the form

$$\mathbf{W}_{\mathcal{A}}^{\mathrm{IC}} = \hat{\mathbf{H}}_{\mathcal{A}} (\hat{\mathbf{H}}_{\mathcal{A}}^\dagger \hat{\mathbf{H}}_{\mathcal{A}})^{-1}, \tag{9}$$

where the columns of $\hat{\mathbf{H}}_{\mathcal{A}}$ are $\hat{\mathbf{h}}_j$, $j \in \mathcal{A}$. At the other end of the spectrum are interference-ignoring (II) multiplexers, for which $\mathbf{W}_{\mathcal{A}}^{\mathrm{II}} = \hat{\mathbf{H}}_{\mathcal{A}}$. For both classes of multiplexer, when there is no quantization, it suffices to replace $\hat{\mathbf{H}}_{\mathcal{A}}$ in $\mathbf{W}_{\mathcal{A}}$ with $\tilde{\mathbf{H}}_{\mathcal{A}}$, the matrix whose columns are $\tilde{\mathbf{h}}_j$, $j \in \mathcal{A}$.

IV. Preliminaries

We begin with a fairly strong notion of optimality of an architecture.

Definition 1: An architecture $\mathcal{S}(P, m)$ is said to be strongly asymptotically optimal (with respect to average throughput) if there exists a sequence of protocols $\mathcal{P}(1), \mathcal{P}(2), \ldots \in \mathcal{S}(P, m)$ such that the corresponding average throughputs $R(1), R(2), \ldots$ of these protocols satisfies

$$\lim_{n\to\infty} \left[R^*(n) - R(n)\right] = 0, \tag{10}$$

[1]With our emphasis on asymptotics, it is sufficient to consider subsets of size m; when not operating in the regime where the asymptotics apply, a generalization that allows the subset size to be a parameter l is more appropriate. See [7] for insights on the choice of l.

where $R^*(n)$ is the best rate achievable by any protocol for the channel and system model of interest.

Note that replacing (10) with the condition

$$\lim_{n\to\infty}\left[\log R^*(n) - \log R(n)\right] = 0 \tag{11}$$

corresponds to a much weaker notion of optimality. To date, most work on asymptotic optimality has focused on this weaker rate-ratio convergence[2], limiting the practical value of the associated results.

To see this weakness, let us define the signal-to-interference+noise ratio $\mathrm{SINR}(n)$ of the protocol via

$$\mathrm{SINR}(n) \triangleq 2^{R(n)/m} - 1. \tag{12}$$

Then weak convergence of rates in the sense of (11) can be obtained even when the SINR gap in dB is asymptotically infinite, i.e., $\mathrm{SINR}^*(n)/\mathrm{SINR}(n) \to \infty$. By contrast, strong convergence of rates in the sense of (11) ensures that the SINR gap in dB is asymptotically zero.

As our main result, we show that the simple, low complexity, decentralized protocol architecture of Section III is strongly asymptotically optimal in the sense of Definition 1 for the channel and system model of Section II.

More specifically, we show that the average throughput achievable by this architecture converges in the sense of (10) to

$$R^*_+(n) = m\log\left(1+\mathrm{SINR}^*(n)\right) + o(1) \tag{13}$$

with

$$\mathrm{SINR}^*(n) = \frac{P\log n}{m^2}, \tag{14}$$

which, as shown in [6], is an asymptotic upper bound on $R^*(n)$, i.e., $\lim_{n\to\infty}\left[R^*_+(n) - R^*(n)\right] \geq 0$.

The average throughput achievable for a given sequence of protocols in our architecture can be expressed in the form

$$R(n) = E\left[R_{\mathbf{H}_{\mathcal{A}}}\right], \tag{15}$$

where the expectation is taken over both the channel realizations and the randomization in the selection of the set $\mathcal{A} \in \mathcal{T}$, and where $R_{\mathbf{H}_{\mathcal{A}}}$ denotes the rate achieved for a particular active set $\mathcal{A}$.

A bound on the rate gap associated with (15) can be readily obtained when there exists, as will be the case in our development, a rate bound $R_-(n)$ such that $R_{\mathbf{H}_{\mathcal{A}}}(n) \geq R_-(n)$ for all $\mathcal{A} \in \mathcal{T}$. In particular, in this case, we may write

$$R(n) \geq (1-p_\emptyset(n))R_-(n), \text{ with } \quad p_\emptyset(n) \triangleq \Pr\{|\mathcal{T}| = 0\},$$

whence

$$R^*(n) - R(n) \leq \left[R^*(n) - R_-(n)\right] + \left[p_\emptyset(n)R_-(n)\right]. \tag{16}$$

Thus to show strong asymptotic optimality, it suffices to show that each of the two terms in brackets in (16) approach zero as $n \to \infty$. We now describe suitable choices for $R_-(n)$ for the particular multiplexers of interest. In the sequel, when there is risk of confusion, we use superscripts $^{\mathrm{II}}$ and $^{\mathrm{IC}}$ to distinguish $R(n)$, $R_-(n)$, $\mathrm{SINR}(n)$, and other quantities for the interference ignoring and cancelling multiplexers, respectively.

[2]We note that strong convergence of random beamforming has recently been shown in [10].

A. Throughput Lower Bounds

Consider first the case of interference-ignoring multiplexers. In this case, for a given active set $\mathcal{A}$ and channel realization $\mathbf{H}_{\mathcal{A}}$, it is straightforward to verify that the achievable sum rate satisfies

$$R^{\mathrm{II}}_{\mathbf{H}_{\mathcal{A}}}(n) = \sum_{j\in\mathcal{A}} \log(1+\mathrm{SINR}^{\mathrm{II}}_j) \tag{17}$$

where

$$\mathrm{SINR}^{\mathrm{II}}_j = \frac{P\|\mathbf{h}_j\|^2\sigma_j^2}{m + P\|\mathbf{h}_j\|^2\|\boldsymbol{\sigma}^{\mathrm{c}}_j\|^2} \tag{18}$$

with

$$\sigma_j = \hat{\mathbf{h}}_j^\dagger\tilde{\mathbf{h}}_j, \quad \text{and} \quad \boldsymbol{\sigma}^{\mathrm{c}}_j = \hat{\mathbf{H}}^\dagger_{\mathcal{A}\setminus j}\tilde{\mathbf{h}}_j. \tag{19}$$

The case for which there is no quantization corresponds to setting $\hat{\mathbf{h}}_j = \tilde{\mathbf{h}}_j$ in (18) and (19), so that $\sigma_j = 1$ and $\boldsymbol{\sigma}^{\mathrm{c}}_j = \tilde{\mathbf{H}}^\dagger_{\mathcal{A}\setminus j}\tilde{\mathbf{h}}_j$.

To obtain a lower bound on $R^{\mathrm{II}}(n)$, we define the following (deterministic) lower bound on $\mathrm{SINR}^{\mathrm{II}}_j$:

$$\mathrm{SINR}^{\mathrm{II}}_-(n) \triangleq \min_{\mathcal{A},j,\mathbf{H}\,:\,|\mathcal{T}|\neq 0,\ \mathcal{A}\in\mathcal{T},\ j\in\mathcal{A}} \mathrm{SINR}^{\mathrm{II}}_j, \tag{20}$$

from which we obtain, via (17) and (15),

$$\frac{R^{\mathrm{II}}(n)}{1-p_\emptyset(n)} \geq \frac{E\left[R^{\mathrm{II}}_{\mathbf{H}_{\mathcal{A}}}\right]}{1-p_\emptyset(n)} \geq m\log(1+\mathrm{SINR}^{\mathrm{II}}_-(n)) \tag{21}$$

for any $\mathcal{A} \in \mathcal{T}$. In turn, via (12), we obtain

$$\mathrm{SINR}^{\mathrm{II}}(n) \geq \left(1+\mathrm{SINR}^{\mathrm{II}}_-(n)\right)^{1-p_\emptyset(n)} - 1. \tag{22}$$

In the absence of quantization there is a corresponding specialization of $\mathrm{SINR}^{\mathrm{II}}_-(n)$.

Considering next the case of interference-cancelling multiplexers, it can be verified that, for a given active set $\mathcal{A}$ and channel realization $\mathbf{H}_{\mathcal{A}}$, the achievable sum rate satisfies

$$R^{\mathrm{IC}}_{\mathbf{H}_{\mathcal{A}}}(n) = \sum_{j\in\mathcal{A}} \log(1+\mathrm{SINR}^{\mathrm{IC}}_j), \tag{23}$$

where, letting $\hat{\boldsymbol{\Phi}}_{\mathcal{A}} = \hat{\mathbf{H}}^\dagger_{\mathcal{A}}\hat{\mathbf{H}}_{\mathcal{A}}$, we have

$$\mathrm{SINR}^{\mathrm{IC}}_j = \frac{P\|\mathbf{h}_j\|^2|\tilde{\sigma}_j|^2}{\mathrm{Tr}(\hat{\boldsymbol{\Phi}}^{-1}_{\mathcal{A}}) + P\|\mathbf{h}_j\|^2\|\tilde{\boldsymbol{\sigma}}^{\mathrm{c}}_j\|^2} \tag{24}$$

with

$$\tilde{\sigma}_j = \frac{\sigma_j - \boldsymbol{\sigma}^{\mathrm{c}\dagger}_j\hat{\boldsymbol{\Phi}}^{-1}_{\mathcal{A}\setminus j}\hat{\boldsymbol{\sigma}}^{\mathrm{c}}_j}{\hat{\sigma}_j - \hat{\boldsymbol{\sigma}}^{\mathrm{c}\dagger}_j\hat{\boldsymbol{\Phi}}^{-1}_{\mathcal{A}\setminus j}\hat{\boldsymbol{\sigma}}^{\mathrm{c}}_j} \tag{25}$$

and

$$\tilde{\boldsymbol{\sigma}}^{\mathrm{c}}_j = \boldsymbol{\sigma}^{\mathrm{c}}_j - \sigma_j\left[\hat{\boldsymbol{\Phi}}_{\mathcal{A}\setminus j} - \hat{\boldsymbol{\sigma}}^{\mathrm{c}}_j\hat{\boldsymbol{\sigma}}^{\mathrm{c}\dagger}_j\right]^{-1}\hat{\boldsymbol{\sigma}}^{\mathrm{c}}_j, \tag{26}$$

which, in turn, are defined in terms of

$$\hat{\sigma}_j = \hat{\mathbf{h}}^\dagger_j\hat{\mathbf{h}}_j = 1 \tag{27}$$

and

$$\hat{\boldsymbol{\sigma}}_j^{\mathrm{c}} = \hat{\mathbf{H}}_{\mathcal{A}\setminus j}^{\dagger}\hat{\mathbf{h}}_j. \tag{28}$$

When there is no quantization we have $\hat{\boldsymbol{\Phi}}_{\mathcal{A}} = \boldsymbol{\Phi}_{\mathcal{A}} = \mathbf{H}_{\mathcal{A}}^{\dagger}\mathbf{H}_{\mathcal{A}}$; $\hat{\sigma}_j = \sigma_j$ so $\tilde{\sigma}_j = 1$; and $\hat{\boldsymbol{\sigma}}_j^{\mathrm{c}} = \boldsymbol{\sigma}_j^{\mathrm{c}}$ so $\tilde{\boldsymbol{\sigma}}_j^{\mathrm{c}} = \mathbf{0}$.

While for the case without quantization a natural bound analogous to (20) is immediate, for the case with quantization it is more convenient to develop an alternative. To this end, we obtain[3]

$$\mathrm{SINR}_j^{\mathrm{IC}} \geq \gamma_j \triangleq \frac{P\|\mathbf{h}_i\|^2\left[|\sigma_j|\tau_j - \sqrt{1-|\sigma_j|^2}\lambda_{\min}\right]_+^2}{\mathrm{Tr}(\hat{\boldsymbol{\Phi}}_{\mathcal{A}}^{-1})\tau_j^2 + P\|\mathbf{h}_j\|^2(1-|\sigma_j|^2)\lambda_{\max}} \tag{29}$$

where $[x]_+ = \max\{0, x\}$ and where $\lambda_{\min}$ and $\lambda_{\max}$ are, respectively, the minimum and maximum eigenvalues of $\hat{\boldsymbol{\Phi}}_{\mathcal{A}\setminus j}$, and where

$$\tau_j = \lambda_{\min} - \|\hat{\boldsymbol{\sigma}}_j^{\mathrm{c}}\|^2. \tag{30}$$

Hence, defining

$$\mathrm{SINR}_{-}^{\mathrm{IC}}(n) \triangleq \min_{\mathcal{A}, j, \mathbf{H}\,:\,|\mathcal{T}|\neq 0,\ \mathcal{A}\in\mathcal{T},\ j\in\mathcal{A}} \gamma_j, \tag{31}$$

which is deterministic, we obtain

$$R^{\mathrm{IC}}(n) \geq (1-p_{\emptyset}(n))\, m\, \log\left(1+\mathrm{SINR}_{-}^{\mathrm{IC}}(n)\right) \tag{32}$$

whence, via (12),

$$\mathrm{SINR}^{\mathrm{IC}}(n) \geq \left(1+\mathrm{SINR}_{-}^{\mathrm{IC}}(n)\right)^{1-p_{\emptyset}(n)} - 1. \tag{33}$$

V. Main Results: Feedback Without Quantization

We now develop the key characteristics of our architecture in the absence of quantization effects.

A. Feedback Requirements

We first characterize the amount of feedback required by the protocol as a function of the parameter settings. For this case, we view $N_\rho = |\mathcal{R}_\rho|$ as a measure of the feedback link capacity requirement. Observe that N_ρ is a binomial random variable with mean $E[N_\rho] = np_\rho$. Since p_ρ is the probability that a user feeds back its channel gain vector, we have, from (5), that $p_\rho = \Gamma(2m, m\rho^-) - \Gamma(2m, m\rho^+)$ with $\Gamma(\cdot,\cdot)$ denoting the incomplete gamma function.

We have the following theorem.

Theorem 1: Let $\rho^+(n) = (\log n)/m$ and $\rho^-(n) = \rho^+(n) - (\log\alpha(n))/m$ where $m\log\log n \leq \log\alpha(n) = o(\log n)$. Then

$$E[N_\rho] = 2m\alpha(n)(1-o(1)) + \mathcal{O}(1/n) \tag{34}$$

From this theorem we see that the choice of $\alpha(n) = e^{m(\rho^+(n)-\rho^-(n))}$ effectively controls the amount of feedback required by the system.

[3]As will become apparent, the appeal of γ_j as a bound is its simple form as $\sigma_j \to 1$.

B. Selection Failure Probability

We next characterize the probability $p_\emptyset$ that the pre-selection phase of the protocol yields no candidate sets.

Theorem 2: Let $\rho^+(n)$ and $\rho^-(n)$ be as in Theorem 1. Then provided $0 \leq \epsilon(n) \leq 1$ we have

$$p_\emptyset(n) \leq e^{-E[N_\rho]\beta(n)/m}, \tag{35}$$

where

$$\log\beta(n) = 2(m-1)^2\log\left(\frac{\epsilon(n)}{2}\right) \tag{36}$$

This theorem characterizes the manner in which successful pre-selection depends on the interference control parameter ϵ and the feedback parameter ρ.

C. Architecture Optimality

Finally, we establish that our architecture is strongly asymptotically throughput optimal.

Theorem 3: Let $\rho^+(n) = (\log n)/m$. If an interference-cancelling multiplexer is used, also let $\rho^-(n) = (\log n)/m - \log\log n$, and $\epsilon(n) = 2/(\log n)^{1/(2(m-1))}$. If an interference-ignoring multiplexor is used, let $\rho^-(n) = (\log n)/m - \left[4(m-1)^2 - 1\right]\log\log n$ and $\epsilon(n) = 2/(\log n)^2$. Then in both cases the protocol sequence $\mathcal{P}_{\epsilon,\rho}(n)$ with average throughputs $R_{\epsilon,\rho}(n)$ and $\mathrm{SINR}_{\epsilon,\rho}(n)$ satisfies

$$R^*(n) - R_{\epsilon,\rho}(n) = \mathcal{O}\left(\frac{\log\log n}{n^2}\right), \tag{37}$$

$$\frac{\mathrm{SINR}^*(n)}{\mathrm{SINR}_{\epsilon,\rho}(n)} - 1 = o(1). \tag{38}$$

Moreover, with this protocol sequence, the feedback link must support, on average,

$$E\left[N_\rho^{\mathrm{IC}}\right] = 2m(\log n)^m(1+o(1)) + \mathcal{O}(1/n) \tag{39}$$

$$E\left[N_\rho^{\mathrm{II}}\right] = 2m(\log n)^{4(m-1)^2+1}(1+o(1)) + \mathcal{O}(1/n) \tag{40}$$

users, depending on which multiplexer is used.

From Theorem 3 we see that while the use of a cruder multiplexer does not incur a penalty in strong throughput optimality, there is a significant price to be paid in terms of the feedback requirement. In particular, in both cases the number of users who must report their channel gains in any scheduling interval is sub-linear, but their sub-linear growth rates are different.

VI. Main Results: Feedback With Quantization

We now generalize our optimality results to the case in which the feedback is quantized.

A. Feedback Requirements

Our result from Section V-A generalizes rather naturally. Since the protocol uses r-bit quantization for each channel gain to be fed back, the total feedback per scheduling interval is $rN_{\rho,\sigma}$ bits, where $N_{\rho,\sigma} = |\mathcal{R}_{\rho,\sigma}|$.

Now N_ρ is similarly a binomial random variable with mean $E[N_{\rho,\sigma}] = np_{\rho,\sigma}$. Since $p_{\rho,\sigma}$ is the probability that a user feeds back its channel gain vector, we have from (4) that

$$p_{\rho,\sigma} = p_\rho p_\sigma \tag{41}$$

where p_ρ is as defined in Section V-A and

$$p_\sigma = \Pr\{|\tilde{\mathbf{h}}_j \hat{\mathbf{h}}_j^\dagger| \geq \sigma\} = 2^r(1-\sigma^2)^{m-1}, \tag{42}$$

with the right-hand equality following from the protocol constraint that $\sigma \geq \mu(\mathcal{C})$, with, as in (3), $\mu(\mathcal{C})$ denoting the coherence of the code. Hence, (41) and (42) imply that the expected aggregate feedback per scheduling interval is proportional to

$$E[N_{\rho,\sigma}] = E[N_\rho]\, 2^r(1-\sigma^2)^{m-1}. \tag{43}$$

B. Selection Failure Probability

We next characterize the probability $p_\emptyset$ that the pre-selection phase of the protocol yields no candidate sets, generalizing our result of Section V-B to the case where there is quantization.

Theorem 4: Let $\rho^+(n)$ and $\rho^-(n)$ be as in Theorem 1. Then for any fixed $\epsilon \geq 0$ we have

$$p_\emptyset(n) \leq e^{-E[N_{\rho,\sigma}]p_{\epsilon|\rho,\sigma}/m} \tag{44}$$

where $E[N_{\rho,\sigma}]$ is as in (43), and where[4]

$$p_{\epsilon|\rho,\sigma} = \frac{k_\epsilon(\mathcal{C}_r)}{\binom{2^r}{m}} \prod_{i=2}^{m} \left(1 - \frac{i-1}{2^r}\right), \tag{45}$$

with $k_\epsilon(\mathcal{C}_r)$ denoting the number of codes of size m with coherence at most ϵ that can be constructed from expurgations of $\mathcal{C}_r$, i.e.,

$$k_\epsilon(\mathcal{C}_r) = \left|\left\{\mathcal{C}_{\log m} \in \mathcal{C}_r \;:\; \mu(\mathcal{C}_{\log m}) \leq \epsilon\right\}\right|. \tag{46}$$

This theorem characterizes the manner in which successful pre-selection depends not only on the feedback parameters (ρ, σ) and the interference control parameter ϵ, but also on the properties of the quantization codebook $\mathcal{C}_r$.

C. Architecture Optimality

Finally, we have that our architecture is also strongly asymptotically throughput optimal when the feedback is quantized.

Theorem 5: Let $\epsilon(n) \equiv 0$, let $\rho^+(n) = (\log n)/m$, $\rho^-(n) = (\log n)/m - (2m-1)/m \cdot \log\log n$, and let $\sigma^2(n) = 1 - 1/\log^2 n$. Furthermore, choose a quantization codebook $\mathcal{C}_r$ such that it contains at least one orthonormal basis, i.e., $k_0(\mathcal{C}_r) \geq 1$. Finally, select the interference-cancelling multiplexer. Then the protocol sequence $\mathcal{P}_{\epsilon,\rho,\sigma}(n)$ with average throughputs $R_{\epsilon,\rho,\sigma}(n)$ and $\mathrm{SINR}_{\epsilon,\rho,\sigma}(n)$ satisfies

$$R^*(n) - R_{\epsilon,\rho,\sigma} = \mathcal{O}\left(\frac{\log\log n}{n^{\psi_r}}\right), \tag{47}$$

$$\frac{\mathrm{SINR}^*(n)}{\mathrm{SINR}_{\epsilon,\rho,\sigma}(n)} - 1 = o(1) \tag{48}$$

[4]Note that $p_{\epsilon|\rho\sigma} = \Pr\{\mathcal{A} \in \mathcal{T}_{\epsilon,\rho,\sigma} \mid \mathcal{A} \subset \mathcal{R}_{\rho,\sigma}\}$.

where $\psi_r = 2^{r+1}p_{\epsilon|\rho,\sigma}$. Moreover, with this protocol sequence, the aggregate rate the feedback link must support, on average, is

$$E[N_{\rho,\sigma}] = 2^{r+1} m \log n(1 + o(1)) + \mathcal{O}(1/n). \tag{49}$$

That one can also get such throughput optimality for the case of interference-ignoring multiplexers follows immediately from the fact that when $\epsilon(n) \equiv 0$ the interference-cancelling and interference-ignoring multiplexers are identical. However, the feedback requirements continue to differ.

For any particular choice of multiplexer, we can also compare the feedback requirement scaling with and without quantization — e.g., (39) and (49) in the case of an interference-cancelling multiplexer. As this case reveals, and as is true more generally, we see that the number of users reporting back their channel gains scales much more slowly when quantization is used. This is because the common quantization is effectively providing sufficient coordination to enable some pre-selection to happen at the receiver.

We also emphasize that the parameter choices in Theorem 5 (and Theorem 3 earlier) are sufficient but not necessary for throughput optimality. And in particular different parameter choices will lead to different tradeoffs between the convergence rate and feedback requirement. However, in the case of quantization, it is worth noting that $\epsilon(n) \to 0$ is necessary.

Finally, it is also worth remarking that an implication of the theorem is that a large codebook (fine quantization) is not required for strong asymptotic throughput optimality — indeed an orthonormal codebook of size m is sufficient. However, the convergence rates do depend on the size and structure of the codebook, and thus quantization codebook design is an important aspect of the overall system design in practice. Such issues are the focus of ongoing work.

REFERENCES

[1] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[2] A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, "Efficient use of side information in multiple- antenna data transmission over fading channels," *IEEE J. Select. Areas Commun.*, vol. 16, no. 8, pp. 1423–1436, Oct. 1998.

[3] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2562–2578, Oct. 2003.

[4] N. Jindal, "MIMO broadcast channels with finite rate feedback," in *Proc. IEEE GLOBECOM*, Nov. 2005.

[5] C. Swannack, E. Uysal-Biyikoglu, and G. W. Wornell, "MIMO broadcast scheduling with limited channel state information," in *Proc. Allerton Conf. Commun., Contr., Computing*, Sep. 2005.

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

[7] C. Swannack, E. Uysal-Biyikoglu, and G. W. Wornell, "Finding NEMO: Near mutually orthogonal sets and applications to MIMO broadcast scheduling," in *Proc. IEEE WIRELESSCOM 2005*, June 2005.

[8] T. Yoo and A. J. Goldsmith, "Optimality of zero-forcing beamforming with multiuser diversity," in *Proc. IEEE ICC*, May 2005.

[9] ——, "Sum-rate optimal multi-antenna downlink beamforming strategy based on clique search," in *Proc. IEEE GLOBECOM*, Nov. 2005.

[10] A. Vakili, A. F. Dana, M. Sharif, and B. Hassibi, "Differentiated rate scheduling for MIMO gaussian broadcast channels," in *Proc. Allerton Conf. Commun., Contr., Computing*, Sep. 2005.