Low Complexity Multiuser Scheduling for Maximizing Throughput in the MIMO Broadcast Channel

Charles Swannack, Elif Uysal-Biyikoglu and Gregory Wornell
Massachusetts Institute of Technology
Dept. of Electrical Engineering and Computer Science
Cambridge, MA 02139
swannack@mit.edu, elif@mit.edu, gww@allegro.mit.edu

Abstract

We consider the general multiuser downlink scheduling problem with n receivers and m transmit antennas, where data from different users can be multiplexed. It is shown that there is a throughput-optimal policy which selects a subset of users to multiplex in each scheduling interval, according to their queue states and current rates achievable by dirty-paper coding. However, the computation of these rates and the selection of the best ordered subset has high complexity in n. It is observed that the simpler multiplexing method of zero-forcing can achieve close to optimal throughput especially when user rate demands are asymmetric. The strong dependence of the objective function with zero-forcing precoding on the queue sizes motivates an algorithm based on ordering users according to queues and making a selection out of a reduced set of users (typically a few multiples of m), which greatly reduces the computational complexity. It is shown that searching beyond this small set is unlikely to improve the objective function.

1 Introduction

We consider the problem of jointly *multiplexing* and *s*cheduling multiple users on the wireless downlink. Figure 1 depicts n data streams to be sent over a multiple-input multiple-output (MIMO) channel where there are m antennas at the transmitter and a single antenna at each of the n receivers.

Multiuser scheduling is the problem of allocating resources (such as power or bandwidth) in order to perform desirably with respect to criteria such as throughput or delay. This problem has attracted great interest in the recent years (*e.g.*, [1, 2, 3, 4, 5, 6]). Most previous studies limit their scope to *time-sharing* schedules, *i.e.* those where only a single user's data is transmitted at any time. The computational complexity of broadcast coding, together with the fact that the optimal coding for the MIMO Broadcast channel was not known until recently [7], has made time-sharing attractive. In fact, transmitting to the user with the best reception is sum-rate optimal (achieves maximum throughput) for a *single-antenna* broadcast channel under *infinite backlogs* and symmetric channels [8]. However, in a multiple-antenna broadcast channel time-division is sub-optimal [9].

Schedules proposed in previous literature also most commonly ignore queuing and randomness in packet arrivals and hence cannot offer stability guarantees. This is true in some scheduling algorithms that aim to satisfy a fairness criteria, such as *proportional-fair* scheduling [2].

A guiding work for incorporating randomness and stability issues has been [10], where the *network capacity region* is defined as the region of stabilizable input data rates, and it is shown that this region is achieved by a *maximum-weight matching* (weights being related to queue sizes). Building on those definitions, [11] considers a broadcast scenario under time division and demonstrates a schedule that achieves the network capacity region. Along similar lines, [12] shows that a throughput-optimal policy is a maximum-weight matching in the form of $\max \sum_i \alpha_i q_i r_i$ where the q_i's are the queue states of users, and the rates r_i are left implicit. Also in the downlink scenario, [13] compares several heuristic scheduling policies such as beamforming to the user with the shortest remaining job versus multiplexing several users. To our knowledge, the maximum-weight matching scheduling policy has first been combined with channel coding and power control explicitly in [14] for the *multiaccess channel*. It was shown that the

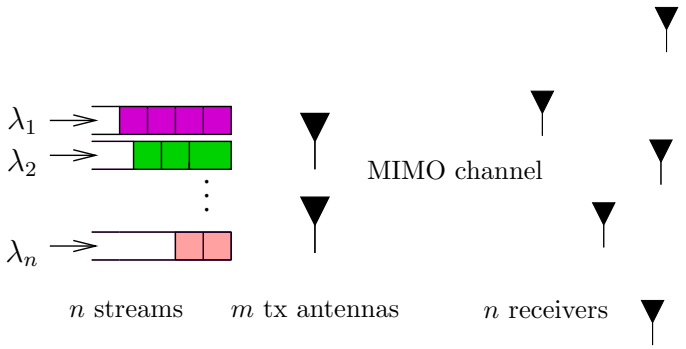


Figure 1: System model: n data streams arriving into separate queues, each stream destined to a single one of the n receivers. There are m transmit antennas, and a single antenna at each receiver. λ_i is the average packet arrival rate for user i.

optimal rate and power control was to assign at each time the longest queue the highest possible rate from the multiaccess channel capacity region.

In this paper we formulate the general broadcast scheduling problem where packet arrivals into the users' queues are random processes. In Section 3 it is established that a throughput-optimal schedule in the downlink channel is one that chooses a group of users and corresponding rates from the *dirty-paper coding* region ([9, 15, 7]) to maximize $\mathbf{q} \cdot \mathbf{r}$. However, this solution has high complexity, not only due to the complexity of obtaining the dirty-paper coding region, but also due to searching for an ordered set of users out of n total users. The complexity of this search is $n!/(n-m)!$, or roughly $\mathrm{O}(n^m)$, which can be formidable as n grows.

Our contribution is to greatly reduce this complexity while keeping the throughput region close to optimal. The complexity reduction is done in two stages: the first is to reduce the complexity by a factor of $m!$ by employing the zero-forcing precoding scheme that is not sensitive to user ordering. Central to this is the observation that a sub-optimal multiplexing scheme, when combined with queue-aware scheduling, can achieve most of the network capacity region. Figure 2 demonstrates this on an example whose details will be clear in the later sections. In the figure, the region of stabilizable input data rates for two users with time-sharing and zero-forcing are shown. Note that when $\lambda_1 \ll \lambda_2$ or $\lambda_2 \ll \lambda_1$ the zero-forcing region comes very close to the outer bound. This observation suggests looking more closely at the dependence of the objective function (specifically, under zero-forcing) on the queue sizes and channel vectors. The second stage of complexity reduction exploits this dependence on queue sizes as well as channel geometry.

The outline of the rest of the paper is as follows. Section 2 presents the MIMO broadcast system model with n users. In Section 3, the network capacity region (the stabilizable region of input rates) is defined and an algorithm that achieves it (and is thus throughput-optimal) is shown. Section 4 derives the scheduling policy under zero-forcing. Zero-forcing is much simpler than optimal multiplexing, however the complexity of searching for a set of m users out of n is still $\binom{n}{m}$. Applying a bound from random graph theory, we show that the probability of finding a quasi-orthogonal set of channels among the first l channels approaches 1 quite fast, which implies that we can restrict attention to a smaller set of users, and not pay a large penalty in the objective function.

2 The MIMO Downlink

We consider the broadcast channel with an m-antenna transmitter and n receivers each having a single receive antenna. Throughout we will let $\mathcal{U}$ be the set of all users and $\mathcal{A}$ be the active set of users (*i.e.* the users that are assigned a non-zero rate at a given time). We will let $\mathbf{x} \in \mathbb{C}^m$ be the transmitted signal vector and $\mathbf{h_i} \in \mathbb{C}^{1\times m}$ be the channel of the ith user. Further, let $\mathbf{H}_{\mathcal{A}}$ be the channel matrix of the

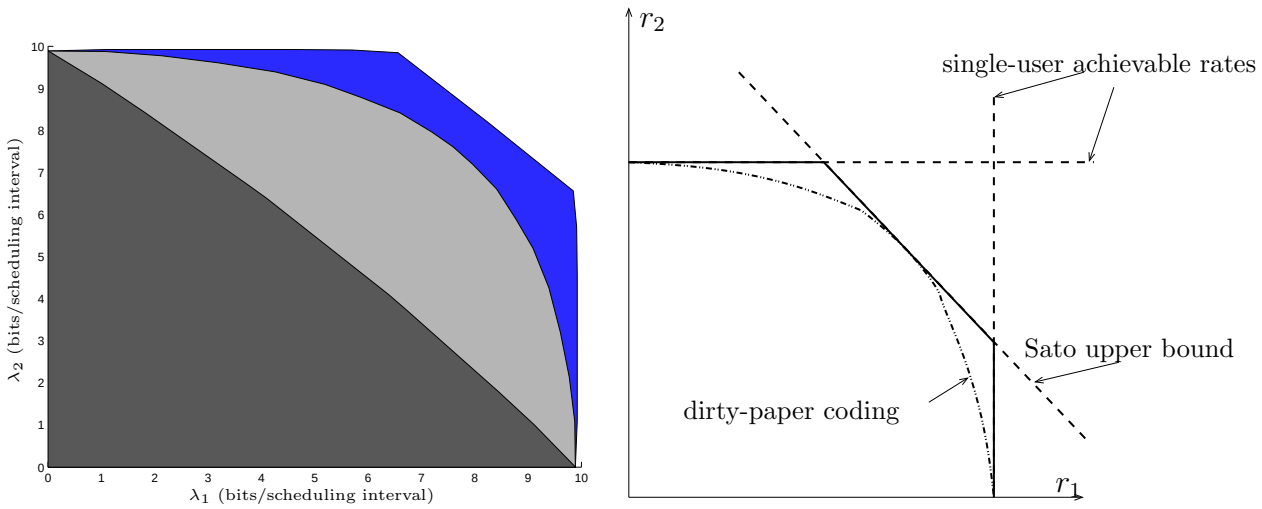


Figure 2: On the left: The regions of stabilizable input rates (λ_1, λ_2) bits/scheduling interval with time-sharing (inner region), zero-forcing and an outer bound. On the right: a description of the outer bound on the optimal rate region used to obtain the outer bound on network capacity. The dirty-paper coding region was not used due to the complexity of computing it.

active users, $\mathbf{x}$ and $\mathbf{y}$ the input and output, respectively. Under the assumption of complex circularly symmetric Gaussian noise we have,

$$\begin{bmatrix} \mathbf{y}_{a_1} \\ \vdots \\ \mathbf{y}_{a_{|\mathcal{A}|}} \end{bmatrix} = \mathbf{H}_{\mathcal{A}}\mathbf{x} + \mathbf{n} \qquad \text{where } \mathbf{H}_{\mathcal{A}} = \begin{bmatrix} \mathbf{h}_{a_1} \\ \vdots \\ \mathbf{h}_{a_{|\mathcal{A}|}} \end{bmatrix}, \text{ and } n_i \sim \mathcal{N}_{\mathbb{C}}(0,1). \tag{1}$$

An upper bound on the input covariance vector, $\mathbf{S}_{xx} = \mathrm{E}[\mathbf{x}\mathbf{x}^\dagger]$, is assumed which corresponds to an total input power constraint of P:

$$\mathrm{Tr}(\mathbf{S}_{xx}) \leq P \tag{2}$$

Recently, it has been shown that the capacity region of the MIMO broadcast channel is achievable by Dirty Paper coding, where users are multiplexed by successively canceling out the interference of other users [7]. The order in which this cancellation occurs influences the users' rates. User i receives the rate $R_{\pi(i)}$ where, π is the permutation that orders the users, given in (3), which also shows the dirty paper coding rate region as the convex hull of the union over the permutations [15]:

$$R_{\pi(i)}(\{\mathbf{S}_{jj}\}) = \log \left| \frac{\mathbf{I} + \mathbf{H}_{\pi(i)} \left(\sum_{j \geq i} \mathbf{S}_{\pi(i)} \right) \mathbf{H}_{\pi(i)}^\dagger}{\mathbf{I} + \mathbf{H}_{\pi(i)} \left(\sum_{j > i} \mathbf{S}_{\pi(i)} \right) \mathbf{H}_{\pi(i)}^\dagger} \right| \quad \mathcal{C}_{\mathrm{DPC}}(\mathbf{H}, P) = co \left(\bigcup_{\substack{\pi \\ \mathrm{Tr}(\mathbf{S_{xx}}) \leq \mathbf{P}}} R(\pi, \{\mathbf{S}_{ii}\}) \right) \tag{3}$$

where $\mathbf{S}_{jj}$ is the input covariance for user j.

3 A Throughput-optimal Scheduling Policy

We assume that the channel vectors $\{\mathbf{h}_i, i = 1, \ldots n\}$ are stationary, ergodic and come from a countable set, such that $\sup \|h_i\| = h_{\max} < \infty$[1]. Packets arrive into buffers i and the arrival processes, $\{A_i(t), i = 1, \ldots n\}$, are modeled as a stationary counting processes with $\lim_{T\to\infty} \frac{A_i(t)}{t} = a_i < \infty$, and $\mathrm{var}(A_i(t+T) - A_i(t)) < \infty$ for $T < \infty$. Packet lengths (in bits) $\{X_i\}$ satisfy $\mathrm{E}(X_i) < \infty$, and $\mathrm{E}(X_i^2) < \infty$, and

[1]This is realistic for practical systems due to quantization.

are independent of arrival instants and of other packet lengths. The arrival processes are independent of each other and the $\{X_i\}$. The *arrival rates* are $\lambda_i = a_i E(X_i)$. The *unfinished work* in queue i at time t is $Q_i(t)$, which is the number of bits waiting to be served in that queue.

Let $\mathcal{C}(P)$ be the region of achievable rate vectors in the broadcast channel under the power constraint given in (2). This is the MIMO broadcast capacity region, achievable by dirty-paper coding over long enough blocklengths, such that if $\mathbf{r} \in \mathcal{C}(P)$, then there exists a power control policy $\mathcal{P}(\mathbf{h})$ and a rate allocation $\mathbf{r}(\mathcal{P}(\mathbf{h}))$ such that $\mathrm{E}_{\mathbf{H}}\mathbf{r}(\mathcal{P}(\mathbf{H})) \geq \mathbf{r}$. For $\mathbf{r}$ outside of $\mathcal{C}(P)$, all power control and rate allocation policies that satisfy (2), $\mathrm{E}_{\mathbf{H}}\mathbf{r}(\mathcal{P}(\mathbf{H})) < \mathbf{r}$.

We adopt the definition of stability in [11] which defines the *overflow function*,

$$g(M) = \limsup_{t\to\infty} \frac{1}{t} \int_0^t 1_{[Q_i(t)>M]} \mathrm{d}\tau$$

and defines queue i to be stable if $g(M) \to 0$ as $M \to \infty$. Accordingly, a rate vector λ is *stabilizable* if there exists a feasible power adaptation and a rate allocation policy under which all queues are stable.

The following is an immediate consequence of Theorem 1 in [16].

Theorem 1. *The broadcast queuing system described above is stabilizable if and only if* $\lambda \in \operatorname{int} \mathcal{C}(P)$

In the following, we present a throughput optimal resource allocation policy for our broadcast system with random packet arrivals. This policy is similar to the throughput optimal resource allocation policy in [11] for a simpler downlink channel, which is extended in [14] to a fading multiaccess channel. The proof follows the technique of [11], also used in [14], which is based on defining an appropriate Lyapunov function and establishing a negative drift on it whenever queue state grows outside a certain compact region.

An additional assumption we shall make is *block-fading*, *i.e.*, channel vectors take independent values at times $t = kT$, for $k \in \mathbb{Z}$, and keep those values in between these time instants. T will assumed to be sufficiently long for error probabilities to be close to zero. Packets will be scheduled every T time units (which we shall call a *s*cheduling interval), *i.e.*, at times $t = kT$, for $k \in \mathbb{Z}$.

Theorem 2. *A throughput optimal resource allocation policy is to assign, at every scheduling interval, feasible rates* $\mathbf{r}(\mathbf{h}, \mathbf{q})$ *and powers* $P(\mathbf{r}(\mathbf{h}, \mathbf{q}))$ *as the solution of the following optimization problem*

$$\max_{\mathbf{r}\in\mathcal{C}_{\mathrm{DPC}}(\mathbf{H},\mathcal{P})} (\mathbf{q}(t) \cdot \mathbf{r}(t)) \tag{4}$$

While the allocation policy (4) does give an explicit optimization problem for optimal throughput, its solution still is very complex computationally. As mentioned, the user selection is of the order $\binom{n}{m}$. Additionally, in the case of dirty-paper coding we must choose an order in which to perform Gram-Schmidt orthogonalization on the selected channels [9], which increases the complexity to to be of order $m!\binom{n}{m}$.

4 Scheduling Under Zero-Forcing Multiplexing

Zero-forcing multiplexing inverts the channel at the transmitter by choosing a transmit vector $\mathbf{x} = \mathbf{H}_{\mathcal{A}}^{+}\mathbf{u}$, where $\mathbf{H}_{\mathcal{A}}^{+}$ is the pseudo-inverse of the channel matrix for the active user set $\mathcal{A}$. We assume throughout that $\mathbf{H}_{\mathcal{A}}$ is non-singular since this occurs with probability 1.

Zero-forcing is attractive in our scenario as a multiplexing technique as it does not incur the user ordering problem and lower coding complexity. In what follows we derive an equivalent expression for (4) which will lead us to derive algorithms of lower complexity.

Let $\mathcal{C}_{ZF}(\mathbf{H}, P)$ be the set of all rate vectors achievable by zero-forcing multiplexing under an power constraint P. Note that the sub-optimality of zero-forcing appears in the power price paid in inverting the channel. It can be shown that the power constraint (2) becomes,

$$\sum_i P_i/b_i \leq P \quad \text{where} \quad b_i = \frac{1}{(W_{\mathcal{A}^{-1}})_{i,i}}, \text{ where } W_{\mathcal{A}} = \mathbf{H}_{\mathcal{A}}\mathbf{H}_{\mathcal{A}}^{\dagger}$$

The b_i's have an important geometric interpretation as noted in [9], as the distance squared of user i's channel from the span of every other users channel in the activation set. This suggests that we pay a large price in power if we have users who are nearly collinear. The dependence on geometry of the optimal set is made more precise through the following theorem and corollary.

Theorem 3. *Let q_i be the queue state for user i. Then under a total power constraint P,*

$$\max_{\mathbf{r}\in\mathcal{C}_{ZF}(\mathbf{H},P)} \mathbf{q}\cdot\mathbf{r} \leq \max_{\substack{\mathcal{A}\subset\mathcal{U}\\ |\mathcal{A}|\leq m}} \left(\sum_{i\in\mathcal{A}} q_i\right)\left(\log\left(1+\frac{P}{\mathrm{Tr}(W_{\mathcal{A}}^{-1})}\right)+D(q_{\mathcal{A}}\|b_{\mathcal{A}})\right) \tag{5}$$

where $q_{\mathcal{A}i}=\frac{q_i}{\sum_{i\in\mathcal{A}} q_i}$, $b_{\mathcal{A}i}=\frac{\frac{1}{b_i}}{\sum_{i\in\mathcal{A}}\frac{1}{b_i}}$ and $D(q_{\mathcal{A}}\|b_{\mathcal{A}})$ is the Kullback Leibler distance.

Proof. Under zero-forcing our problem becomes, $f(q,\mathcal{A})=\max_{\sum_i P_i/b_i\leq P}\sum_i q_i\log(1+P_i)$. Using convexity, writing the Lagrangian and taking the derivative yields the necessary and sufficient condition

$$P_i=\lambda q_i b_i-1 \qquad \text{where} \qquad \sum_i\left(q_i\lambda-\frac{1}{b_i}\right)_+=P \tag{6}$$

Now, suppose every user gets a strictly positive power. Then, noting that $\lambda=\frac{P+\sum_i\frac{1}{b_i}}{\sum_{i\in\mathcal{A}} q_i}$ yields

$$f(q,\mathcal{A})=\sum_i q_i\log(1+P_i)=\left(\sum_i q_i\right)\log(\lambda)+\sum_i q_i\log(b_i\ q_i) \tag{7}$$

Now using the cofactor expansion for the inverse we have, $b_i=\frac{\det(\mathbf{H}_{\mathcal{A}}\mathbf{H}_{\mathcal{A}}^{\dagger})}{\det(M_{ii}(\mathbf{H}_{\mathcal{A}}\mathbf{H}_{\mathcal{A}}^{\dagger}))}=\frac{A(\mathcal{A})}{M_{i,i}(\mathcal{A})}$. Returning to (7) we have

$$\begin{aligned} f(q,\mathcal{A}) &\leq \left(\sum_{i\in\mathcal{A}} q_i\right)\log\left(\frac{P+\frac{1}{A(\mathcal{A})}\sum_i M_{i,i}(\mathcal{A})}{\sum_{i\in\mathcal{A}} q_i}\right)-\sum_i q_i\log\left(\frac{M_{i,i}(\mathcal{A})}{q_i A(\mathcal{A})}\right)\\ &= \left(\sum_{i\in\mathcal{A}} q_i\right)\left(\log\left(\frac{PA(\mathcal{A})}{\sum_i M_{i,i}(\mathcal{A})}+1\right)+\sum_{i\in\mathcal{A}}\tilde{q}_i\log\left(\frac{\tilde{q}_i}{\tilde{m}_i}\right)\right) \qquad (8)\\ &= \left(\sum_{i\in\mathcal{A}} q_i\right)\left(\log\left(\frac{P}{\sum_{i=1}^m\frac{1}{b_i}}+1\right)+D(q_{\mathcal{A}}\|m_{\mathcal{A}})\right) \end{aligned}$$

where $m_{\mathcal{A}i}=\frac{M_{i,i}(\mathcal{A})}{\sum_{i\in\mathcal{A}} M_{i,i}(\mathcal{A})}$

□

Corollary 1. *Let $\mathcal{U}^+(\mathbf{H},P,\mathbf{q})$ be the set of all user sets such that the optimal power allocation gives each user of $\mathcal{A}\in\mathcal{U}^+(\mathbf{H},P,\mathbf{q})$ a strictly positive power (i.e. $P_i>0$ in (6)). Then,*

$$\max_{\mathbf{r}\in\mathcal{C}_{ZF}(\mathbf{H},P)}\mathbf{q}\cdot\mathbf{r}=\max_{\substack{\mathcal{A}\in\mathcal{U}^+(\mathbf{H},P,\mathbf{q})\\ |\mathcal{A}|\leq m}}\left(\sum_{i\in\mathcal{A}} q_i\right)\left(\log\left(1+\frac{P}{\mathrm{Tr}(W_{\mathcal{A}}^{-1})}\right)+D(q_{\mathcal{A}}\|b_{\mathcal{A}})\right) \tag{9}$$

Furthermore,

$$\max_{\substack{\mathcal{A}\in\mathcal{U}^+(\mathbf{H},P,\mathbf{q})\\ |\mathcal{A}|\leq m}}\left(\sum_{i\in\mathcal{A}} q_i\right)\left(-H(\mathbf{q})+q_{\mathcal{A}}^*\log|W_{\mathcal{A}}|+\hat{f}(\mathcal{A})\right)\geq\max_{\substack{\mathcal{A}\in\mathcal{U}^+(\mathbf{H},P,\mathbf{q})\\ |\mathcal{A}|\leq m}}\mathbf{q}\cdot\mathbf{r}\geq\max_{\substack{\mathcal{A}\in\mathcal{U}^+(\mathbf{H},P,\mathbf{q})\\ |\mathcal{A}|\leq m}}\left(\sum_{i\in\mathcal{A}} q_i\right)\tilde{f}(\mathcal{A}) \tag{10}$$

where $\tilde{f}(\mathcal{A})=\log\left(1+\frac{P}{\mathrm{Tr}(W_{\mathcal{A}}^{-1})}\right)$ and $\hat{f}(\mathcal{A})=\tilde{f}(\mathcal{A})+\log\left(\mathrm{Tr}\left(W_{\mathcal{A}}^{-1}\right)\right)$ and $q_{\mathcal{A}}^*=\frac{1}{\sum_{i\in\mathcal{A}} q_i}\arg\max_{i\in\mathcal{A}} q_i$

Proof. The lower bound follows from that fact that $D(q_{\mathcal{A}}\|b_{\mathcal{A}})$ is greater than or equal to zero [17]. To arrive at the upper bound note that,

$$\begin{aligned} D(q_{\mathcal{A}}\|b_{\mathcal{A}}) &= \sum_{i\in\mathcal{A}} \tilde{q}_i \log\left(\frac{\tilde{q}_i}{\tilde{b}_i}\right) \\ &= -H(\mathbf{q}) + \log\left(\mathrm{Tr}\left(W_{\mathcal{A}}^{-1}\right)\right) + \sum_{i\in\mathcal{A}} \tilde{q}_i \log\left(b_i\right) \\ &\leq -H(\mathbf{q}) + \log\left(\mathrm{Tr}\left(W_{\mathcal{A}}^{-1}\right)\right) + q_{\mathcal{A}}^{*} \log\left(\prod_{i=1}^{l} \frac{1}{w_{ii}}\right) \text{ where } w_{ii} = \left(W_{\mathcal{A}}^{-1}\right)_{ii} \\ &\leq -H(\mathbf{q}) + \log\left(\mathrm{Tr}\left(W_{\mathcal{A}}^{-1}\right)\right) + q_{\mathcal{A}}^{*} \log\left(|W_{\mathcal{A}}|\right) \text{ (by Hadamard's Inequality)} \end{aligned}$$

□

The preceding corollary provides significant insight into user selection for zero-forcing multiplexing. In particular, (9) may be broken up into three main components (see Figure 3). First, examining the objective function (9) one can see that it is almost linear in the sum of the queue states of the active users. The only non-linearity comes from the distance term. We call this leading linear term the *queuing gain* since it captures the significance of the magnitude of the queue state on the objective function. Further, we define the logarithmic term as the *geometry gain* since it captures the effects the geometry of the channels have on the optimal solution (recall that the b_i can be interpreted as the square of the distance of each user's channel vector from the subspace spanned by the other users' channels [9]). We note that these two terms are "bulk" properties of the vectors of queue states and channel states and are decoupled from one another. The inter-dependence of these two vectors is captured in the *pairing gain*. This term reflects the reward one gets for choosing a set of users whose queue state is matched well with its channel.

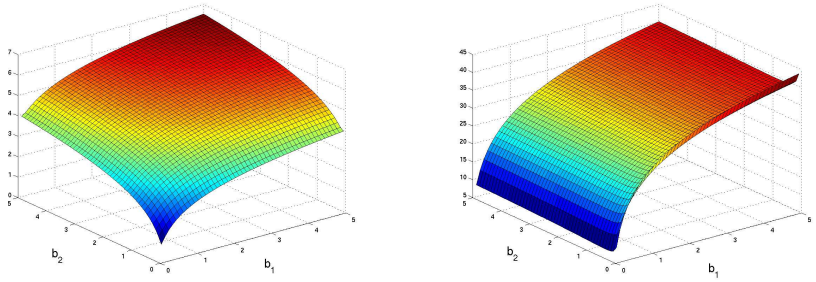
$$\underbrace{\left(\sum_{i\in\mathcal{A}} q_i\right)}_{\text{queuing gain}} \left(\underbrace{\log\left(1 + \frac{P}{\mathrm{Tr}(W_{\mathcal{A}}^{-1})}\right)}_{\text{geometry gain}} + \underbrace{D(q_{\mathcal{A}}\|b_{\mathcal{A}})}_{\text{pairing gain}} \right)$$

Figure 3: The decomposition of (5) into its various components

It is important to note that the RHS of (10) *is* linear in the magnitude of the queue state and is approximately equal to the LHS at high SNR. This suggests that when searching for an optimal solution among a large set of users, ignoring users with queue states an order of magnitude smaller than the largest queue will, with high probability, not incur a significant penalty. For example, we can see from Figure 4 that for $m = 2$ the user whose queue is an order of magnitude smaller must have a channel vector whose component orthogonal to the other user's channel is exponentially larger. This can also explain why in Figure 2 for $\lambda_1 \gg \lambda_2$ zero forcing precoding approaches the outer bound. The dominance of the first user's queue size on rate allocation effectively turns this into a single user channel. Therefore, if there is a dominant class of users we can restrict our search to this set. Otherwise, the queues are approximately balanced, and in this case we can approximate the optimal solution by optimizing $\tilde{f}(\mathcal{A})$. We note that this term is only a function of the geometry of the channel states. The dominance of the geometry motivates selecting a set of users with nearly orthogonal channels.

4.1 Selecting Nearly Orthogonal Sets

For a given set of channel magnitudes and corresponding queue states, the weighted sum rate is maximized by an orthogonal set. Using our definitions, such a set has the largest conditional "geometry



(a) (b)

Figure 4: The objective function for the optimization problem (5) where $m = 2$ as a function of b_i. For (a) $q_1 = 1$, $q_2 = 1$ and (b) $q_1 = 10$, $q_2 = 1$

gain". In addition, in this case since the channel matrix is orthogonal, zero-forcing incurs no power penalty for inverting the channel and thus can approach the dirty-paper coding rate.

Of course, for finite n, the probability of existence of an orthogonal set is zero. In this section we address the probability of existence of a "nearly" orthogonal set. To be precise, define two vectors $\mathbf{h_1}$ and $\mathbf{h_2}$ to be ϵ-orthogonal if $|\mathbf{h_1}\mathbf{h_2^\dagger}| \leq \epsilon$ and define $a_{ij}(\epsilon, \delta) = \mathbf{1}_{\{|\mathbf{h_1}\mathbf{h_2^\dagger}| \leq \epsilon \cap |\mathbf{h_1}| > \delta \cap |\mathbf{h_2}| > \delta\}}$ as the indicator function of two vectors being ϵ-orthogonal with magnitudes of at least δ. Note that this is a symmetric relationship. We let $F_m(\epsilon, \delta)$ be the event that there exists a set of m receivers who are mutually ϵ-orthogonal with magnitudes at least δ. More formally, $F_m(\epsilon, \delta) = \{\exists \mathcal{A}, |\mathcal{A}| = m : \ |\mathbf{h_i}\mathbf{h_j^\dagger}| < \epsilon \ \forall i, j \in \mathcal{A} \ \cap \ |\mathbf{h_i}| > \delta \ \forall i \in \mathcal{A}\}$

We will obtain upper and lower bounds on the probability of existence of an ϵ-orthogonal set using results from random graph theory[2]. The following definitions are in order.

Let $\mathbf{A}$ be the matrix with entries a_{ij} and consider the undirected graph $G(\mathbf{A})$ with adjacency matrix $\mathbf{A}$. Also define $X_E(G)$ to be the number of isomorphic copies of the graph E in G (we adopt the terminology of [18].) Now, in order for the event $F_m(\epsilon, \delta)$ to occur we must have $a_{ij} = 1 \ \forall \ i, j \in \mathcal{I}$ where $\mathcal{I}$ is some index set of cardinality m. More precisely we have the following.

Lemma 1. $\Pr(F_m(\epsilon, \delta)) = \Pr(X_{K_m}(G(\mathbf{A})) \geq 1)$ *where* K_n *is the complete graph on* n *vertices.*

Let $G(n, q)$ to be the random graph on n vertices with edges chosen with probability q and define

$$p = p_\epsilon(m, \delta) = \Pr\left(|\mathbf{h_1}\mathbf{h_2^\dagger}| \leq \epsilon \ \cap \ |\mathbf{h_1}| > \delta \ \cap \ |\mathbf{h_2}| > \delta\right)$$

as the probability that any two users are ϵ-orthogonal and

$$\tilde{p} = \tilde{p}_\epsilon(m, \delta) = \Pr\Big(|\mathbf{h_m}\mathbf{h_i^\dagger}| \leq \epsilon \ i = 1 \ldots m-1 \ \cap \ |\mathbf{h_m}| > \delta \mid |\mathbf{h_j}\mathbf{h_i^\dagger}| \leq \epsilon \ \forall \ 0 < i < j < m \\ \cap \ |\mathbf{h_i}| > \delta \ i = 1 \ldots m-1 \ \Big)$$

be the probability that a given vector $\mathbf{h}$ is ϵ-orthogonal to a set of $m-1$ who are mutually ϵ-orthogonal.

Theorem 4. *Let* $\{\mathbf{h_j}\}$ *be a set of* n *independent random vectors where* $\mathbf{h_j} \in \mathbb{C}^{m \times 1}$ *and* $\mathbf{h_j} \sim \mathcal{N}_\mathbb{C}(0, \mathbf{I})$. *Then,*

$$1 - e^{-ck_m(p)} \leq \Pr\left(F_m(\epsilon, \delta)\right) \leq 1 - e^{-\frac{k_m(\tilde{p})}{(1-\tilde{p})}} \quad \textit{where} \quad k_m(q) = \binom{n}{m} q^{\frac{m(m-1)}{2}}$$

Proof. See Appendix A.

[2]We note that these bounds may be derived in an alternative fashion, but this theory offers a more compact proof

4.2 Discussion

The probability of finding ϵ-orthogonal set increases to 1 as we search in a larger and larger set of users. As indicated in Figure 5 (a), this probability curve has a knee at some value l where it has made most of its increase, and looking at a larger set of users than l will not improve the objective function significantly. The significance of this is that this as a low complexity algorithm in a practical system. For that, of course, it would be useful to be able to determine the l at which the phase transition occurs.

Figure 5 (b) depicts the throughput region for two users, obtained for $m = 4$ transmit antennas and $n = 24$ users which form two separate rate classes such that 12 of the users have independent arrival processes at rate λ_1, and the remaining 12 have independent arrival processes at rate λ_2. The arrival rates given belong to one user from each class. The separate curves correspond to increasing $l = km$, for $k = 3, 4, 5, 6$. The outermost curve, which corresponds to $k = 6$, considered the whole set of users in the selection, hence is the maximum rate curve achievable in this scenario with zero-forcing. Note that the curves become close faster as we increase l. Also note that due to simulation time constraints, we restricted attention to a small number of antennas m and a small n; for larger m and n we expect the curve at $l = km$ to be closer to the curve $l = n$ (the optimal) for smaller values of k.

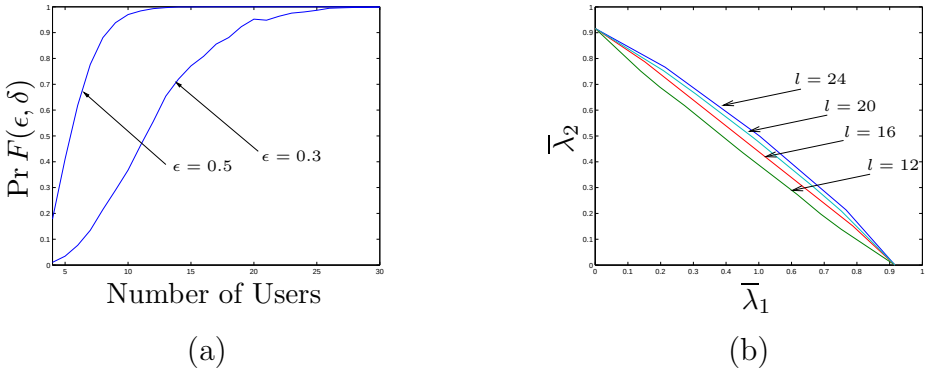


Figure 5: (a) The empirical probability of finding an orthogonal set for $\epsilon = 0.3$ and $\epsilon = 0.5$ with $\delta = 0$ (b) The throughput region for $m = 4$ transmit antennas and $n = 24$ users, which form two separate rate classes under various scheduling policies.

5 Conclusion

We have addressed the joint multiplexing and scheduling problem in the multiantenna broadcast channel. We first established that a throughput optimal policy is one that picks a subset of users in each scheduling interval to maximize $\mathbf{q} \cdot \mathbf{r}$, where q is the queue state vector, and r is the rate vector from the dirty-paper coding region for the given channel state. However, the computational complexity of the set selection is excessive. We observe that a sub-optimal multiplexing strategy used alongside an optimal scheduling strategy can achieves most of the network capacity region. In particular, with zero-forcing multiplexing one needs only search over unordered sets, which reduces complexity by a factor of $m!$ (where m is the number of transmit antennas.) By examining the $\mathbf{q} \cdot \mathbf{r}$ function under zero-forcing, we see that the near-optimal performance at asymmetric input rates is due to the strong dominance of a large queue state over a smaller one. This suggests a practical low-complexity scheduling strategy simply orders users according to decreasing queue size and simply disregards those that have a queue size an order of magnitude or more smaller than the largest. Of course, when queues are balanced (as in the case of symmetric arrival rates), such an approach does not have an important complexity advantage. The key observation is to notice that when queues are approximately balanced, the objective function is dominated by a function that is related to channel geometry, and that the optimal geometry can be found among a quite small set of users. More precisely, the probability of existence of an almost

orthogonal set of channels increases to 1 rapidly as the number users increases. This implies that one can, with high probability, obtain close to maximum throughput with a search of complexity $\binom{l}{m}$, where l is typically a few multiples of m.

The work in this paper suggests the existence of low complexity joint multiplexing/scheduling algorithms that are expected to have good throughput. It is the subject of future work to make precise the complexity vs. throughput tradeoff and derive specific algorithms to achieve certain points on that tradeoff.

A Proof of Theorem 4

In order to prove theorem 4 we require the following lemma and proposition.

Lemma 2.

$$\Pr\left(X_{K_m}(G(n,\tilde{p})) \geq 1\right) \leq \Pr\left(X_{K_m}(G(\mathbf{A})) \geq 1\right) \leq \Pr\left(X_{K_m}(G(n,p)) \geq 1\right)$$

Proof. To prove lemma 2 it is sufficient to show

$$\Pr\left(K_m \subset G(n,\tilde{p})\right) \leq \Pr\left(K_m \subset G(\mathbf{A})\right) \leq \Pr\left(K_m \subset G(n,p)\right)$$

i.e. that the probability of any copy of K_m in $G(\mathbf{A})$ is greater than (resp. less than) a copy of K_m in $G(n,\tilde{p})$ (resp. $G(n,p)$). Now, we note that the edges of $G(\mathbf{A})$ are independent if they do not share a common vertex by the independence of the $\mathbf{h}_i$ but are not independent of all other edges. In general,

$$\begin{aligned}
\Pr\left(K_m \subset G(\mathbf{A})\right) &= \Pr\left(\{\cap_{i<j}\{a_{ij}=1\}\right) \\
&= \prod_{i=1}^{m}\prod_{j=1+1}^{m} \Pr\left(a_{ij}=1|\{\cap_{k=1}^{i-1}\cap_{l=k+1}^{m} a_{ij}=1\} \cap \{\cap_{q=i+1}^{j-1} a_{ij}=1\right) \\
&\leq \prod_{i<j} \Pr\left(a_{ij}=1\right) = p^{m(m-1)/2} = \Pr\left(K_m \subset G(n,p)\right)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\Pr\left(K_m \subset G(\mathbf{A})\right) &= \Pr\left(\{\cap_{i<j}\{a_{ij}=1\}\right) \\
&= \prod_{i<j}^{m} \Pr\left(a_{ij}=1|\{\cap_{k=1}^{i-1}\cap_{l=k+1}^{m} a_{ij}=1\} \cap \{\cap_{q=i+1}^{j-1} a_{ij}=1\}\right) \qquad (11) \\
&\geq \prod_{i<j} \Pr\left(a_{ij}=1|\cap_{\substack{k<j \\ k\neq i}}\{a_{ij}=1\}\right) \\
&= \hat{p}^{m(m-1)/2} = \Pr\left(K_m \subset G(n,\hat{p})\right) \qquad (12)
\end{aligned}$$

■

Proposition 1. *[19] Let F be a fixed non-empty graph, then*

$$e^{-\Phi_F/(1-p)} \leq \Pr\left(F \not\subset G(n,p)\right) \leq e^{-c\Phi_F}$$

where

$$\Phi_F = \min\{\mathrm{E}\{X_H(G(n,p))\} : H \subset F, e(H) > 0\}$$

The result follows from setting $F = K_m$ in the preceding lemma and noting that $\min\{\mathrm{E}\{X_H(G(n,p))\} : H \subset K_m, e(H) > 0\} = \mathrm{E}\{\mathrm{X_{K_m}(G(n,p))}\} = \binom{n}{m} p^{\frac{m(m-1)}{2}}$

■

References

[1] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, 2001.

[2] D. Tse, "Multiuser diversity in wireless networks," presentation, Stanford University, April 2001.

[3] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, June 2002.

[4] X. Liu, E. Chong, and N. Shroff, "Joint scheduling and power-allocation for interference management in wireless networks," in *Proc. 56th IEEE Vehicular Technology Conference*, vol. 3, Vancouver, British Columbia, September 2002, pp. 1892–1896.

[5] M. Lopez, "Multiplexing, scheduling and multicasting strategies for antenna arrays in wireless networks," Ph.D. dissertation, MIT, Cambridge, Massachusetts, 2002.

[6] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *INFOCOM 2003. Proc. of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies*, San Francisco, 2003.

[7] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the gaussian mimo broadcast channel," in *Proc. CISS*, Princeton University, March 2004.

[8] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels: Part I: Ergodic capacity," *IEEE Trans. Inform. Theory*, vol. 47, no. 3, pp. 1083–1102, March 2001.

[9] 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.

[10] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Automat. Contr.*, vol. 37, no. 12, pp. 1936–1948, December 1992.

[11] M. Neely, E. Modiano, and C. Rohrs, "Power allocation and routing in multibeam satellites with time-varying channels," *IEEE/ACM Trans. Networking*, vol. 11, no. 1, pp. 138–152, February 2003.

[12] H. Vishwanathan and K. Kumaran, "Rate scheduling in multiple antenna downlink wireless systems," in *Proc. 39th Annual Allerton Conf. on Communication, Control, and Computing*, Monticello, Illinois, 2001.

[13] M. Airy, S. Shakkottai, and R. Heath, Jr., "Limiting queuing models for scheduling in multi-user mimo wireless systems," in *Proc. of the 2nd IASTED Conf. on Communications, Internet & Info Technology*, Scottsdale, Arizona, November 2003.

[14] E. Yeh and A. Cohen, "Maximum throughput and minimum delay in fading multiaccess communications," to be published.

[15] S. Vishwanath, N. Jindal, and A. Goldsmith, "On the duality of gaussian multiple-access and broadcast channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 5, pp. 768–783, May 2004.

[16] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," in *INFOCOM 2003. Proc. of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 1, San Francisco, 2003, pp. 745–755.

[17] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons Inc., 1997.

[18] B. Bollobás, *Modern Graph Theory*, ser. Graduate Texts in Mathematics. New York: Springer-Verlag, 1998, no. 184.

[19] ——, *Random Graphs*, 2nd ed., ser. Cambridge studies in advanced mathematics. Cambridge: Cambridge University Press, 2001, no. 73.