MIMO-OFDMA Rate Allocation and Beamformer Design Using a Multi-Access Channel Framework

Ali Khanafer
Univ. of Illinois at Urbana-Champaign
khanafe2@illinois.edu

Teng Joon Lim
National Univ. of Singapore
eleltj@nus.edu.sg

Roya Doostnejad
Univ. of Toronto
r.doostnejad@utoronto.ca

Taiwen Tang
BLiNQ Networks Inc.
eric.tang@blinqnetworks.com

Abstract—This paper tackles the downlink user scheduling and transmit beamforming problems in MIMO-OFDMA by extending a recent algorithm that maximizes the weighted sum rate (WSR) to all users in MIMO flat fading channels. The proposed method has a complexity that is proportional to the number of OFDMA subcarriers, which makes it practically attractive. Having assigned users to each subcarrier and designed beamformers for each user in each subcarrier, it remains to find the best, in terms of rate maximization, adaptive modulation and coding (AMC) mode to use for each data stream. The latter problem is solved in the second half of the paper through viewing the channel from the base station to the \( k \)-th receiver as a multiple access channel (MAC) with \( N_k \) “users”, where \( N_k \) is the number of antennas at receiver \( k \). The proposed method maps the available AMC modes to the space of allowed theoretical rates, using the signal-to-noise ratio (SNR) gap to capacity concept, and selects the operating point with the largest sum-rate.

I. INTRODUCTION

In order to serve a high throughput network with a large number of high-mobility users, system designers have increasingly turned to multi-antenna multi-carrier transceivers, also known as MIMO-OFDMA systems. Well-known examples of such systems include WiMax and LTE for the next generation of commercial cellular communications [1].

Because OFDMA transforms a frequency-selective channel into \( N \) orthogonal (or parallel) frequency-nonselective channels, one can always treat a MIMO-OFDMA system as if it were made up of \( N \) independent MIMO flat-fading channels. This simplistic perspective allows for the straightforward extension of MIMO methods (e.g. for scheduling or precoding) to MIMO-OFDMA, by designing on a per-subcarrier basis with equal transmitted power on each subcarrier. However, such a method is sub-optimal because user scheduling must fundamentally be performed across all subcarriers, since each user has access to the entire OFDMA frequency band. The main hurdle to joint precoding and scheduling over the entire OFDMA band, in which the frequency dimension is written into the signal model, is complexity that scales faster than linearly with the number of subcarriers.

In this paper, we focus on maximizing weighted sum rate (WSR) because (a) the weights can be used to prioritize users, e.g. giving more importance to a user with streaming traffic, and (b) a recent WSR-maximizing algorithm [2] for the MIMO-BC can be extended to MIMO-OFDMA with a resulting complexity that scales only linearly with the number of OFDM sub-carriers. Point (b) is a unique feature of the algorithm in [2], as many other MIMO-BC precoding techniques in the literature must solve geometric programs within nested iterations [3], [4], [5], or signomial programs [6], leading to very much higher complexity even in the single-carrier case. With \( N \) subcarriers in OFDMA, the optimization variables have \( N \) times the dimensionality of the single-carrier case, and hence non-linear complexity scaling is not tolerable.

Here, we extend the WSRBF-WMMSE algorithm of [2] to MIMO-OFDMA systems. Its low complexity is due to its reliance on closed-form expressions which make the computations very efficient – we will refer to this algorithm as WW-OFDMA. We will show that WW-OFDMA can select users and allocate subcarriers to users automatically. A critical point we will highlight is that the complexity of WW-OFDMA scales linearly with the number of subcarriers as opposed to the worst case polynomial-time complexity in the number of parameters of the optimization problem incurred by algorithms requiring a GP to be solved.

Another problem tackled in this paper is rate allocation for the streams of each user. With all of the WSR maximizing algorithms, a user with \( N_k \) receive antennas will generally be allocated a beamforming matrix of rank \( N_k \), i.e., the BS will transmit \( N_k \) independent data streams to that user. The rate to be transmitted to the \( k \)-th user is \( \tilde{R}_k \), which has to be split among \( N_k \) streams. This rate-splitting problem does not have a unique solution, because the channel between the BS and user \( k \), with beamforming vectors fixed by one of the above algorithms, is effectively a multiple access channel (MAC) with \( N_k \) “users” (as shown later) and any point on the maximum sum-rate surface (the ~45 degree face for the two-user MAC) achieves the desired rate \( \tilde{R}_k \) with optimal decoding and ideal Gaussian codes.

Using this MAC equivalence perspective, and the signal-to-noise ratio (SNR) gap to capacity concept, we go on to design an adaptive modulation and coding (AMC) scheme for each stream of each user, that splits the data rate for each stream among \( N_k \) streams and results in performance better than a given desired bit error probability \( P_b \). To do this, we assume that a quadrature amplitude modulation (QAM) format is used, and that the coding gain of the error control code employed is known. Although we limit our analysis here to beamform-
ers designed by the WSRBF-WMMSE algorithm due to its simplicity, our technique is general and can be applied to beamformers designed using other approaches. Moreover, the MAC equivalence we outline here can handle different receiver structures making it a universal method for per stream rate allocation. Alternative to the SNR gap to capacity, one could use the optimal approach devised by Lozano et al. in [7] which is based on maximizing mutual information; this is a subject of future work.

The rest of this paper is organized as follows. In Section II we present our WSR maximization algorithm for MIMO-OFDMA systems and discuss its computational complexity. In Section III we describe the AMC problem and suggest a stream rate allocation algorithm. We evaluate the performance of the presented algorithms in Section IV and conclude the paper in Section V.

II. RESOURCE ALLOCATION AND BEAMFORMING DESIGN

A. Signal Model

Consider the downlink of a MIMO-OFDMA wireless system where the BS employs $M$ transmit antennas and broadcasts to $K$ users over $N$ subcarriers. User $k$ has $N_k$ antennas, and the total number of receive antennas is $N_r = \sum_{k=1}^{K} N_k$. Assuming that the OFDM cyclic prefix is sufficiently large, there will be no inter-symbol interference, and hence we can focus on one OFDM symbol without loss of generality. Let $\mathbf{d}_k[n] \in \mathbb{C}^{N_k \times 1}$ be the symbol vector intended for the $k$-th user on the $n$-th subcarrier$^1$ with $E[\mathbf{d}_k[n]\mathbf{d}_k[n]^H] = \mathbf{I}$. Let the transmit beamformer to the $k$-th user on the $n$-th subcarrier be $\mathbf{B}_k[n] \in \mathbb{C}^{M \times N_k}$. The transmitted signal on the $n$-th subcarrier is thus given by $\mathbf{x}[n] = \sum_{k=1}^{K} \mathbf{B}_k[n]\mathbf{d}_k[n]$. Let $\mathbf{H}_k[n] \in \mathbb{C}^{N_k \times M}$ be the channel from the BS to the $k$-th user on the $n$-th subcarrier where the $(i,j)$-th element of $\mathbf{H}_k[n]$ is the channel gain from the $j$-th transmit antenna to the $i$-th receive antenna. The received signal at the $k$-th user on the $n$-th subcarrier would be

$$\mathbf{y}_k[n] = \mathbf{H}_k[n]\mathbf{x}[n] + \mathbf{w}_k[n]$$

where $\mathbf{w}_k[n] \in \mathbb{C}^{N_k \times 1}$ is a noise vector with independent zero-mean circularly symmetric Gaussian (ZMCSCG) elements and covariance matrix $\mathbf{I}$.

In [2], a linear pre-processing method was described for selecting a subset of users $\mathcal{K}$ to serve among all available users, and for finding the beamforming matrices for all users in $\mathcal{K}$, in a flat fading MIMO-BC. The criterion used in this design is the weighted sum rate (WSR) $\sum_{k \in \mathcal{K}} \mu_k R_k$ subject to a total power constraint $P_{\text{max}}$ where $R_k$ is the rate assigned to user $k$, and $\mu_k$ is a pre-determined value that indicates the relative importance of user $k$. The channels to all users are assumed to be perfectly known at the BS. In the following section, we extend this so-called WSRBF-WMMSE algorithm to MIMO-OFDMA.

B. WW-OFDMA

There are two possible approaches towards applying WSRBF-WMMSE in MIMO-OFDMA systems (1) perform resource allocation jointly over all subcarriers by extending the definition of the beamformers and channel matrices of each user to contain the frequency dimension; (2) run WSRBF-WMMSE for each subcarrier independently subject to a per subcarrier power constraint (uniform power allocation across subcarriers). Clearly, Approach 1 is the optimal approach to allocate subcarriers to different users or, equivalently, perform user selection across subcarriers. On the other hand, Approach 2 appears to be simpler as no joint processing is required. However, we will show that both approaches have the same complexity.

To adopt Approach 1, we need to re-write the received signal (1) to span all subcarriers. This is done by extending the definitions of the channels and beamforming matrices to include both the spatial and frequency dimensions as follows:

$$\mathbf{y}_k = [\mathbf{y}_k[1]^T, \mathbf{y}_k[2]^T, \ldots, \mathbf{y}_k[N]^T]^T$$

$$= \mathbf{H}_k \sum_{k=1}^{K} \mathbf{B}_k \mathbf{d}_k + \mathbf{w}_k, \quad (2)$$

where $\mathbf{H}_k = \text{blkdiag}\{\mathbf{H}_k[1], \mathbf{H}_k[2], \ldots, \mathbf{H}_k[N]\}$, $\mathbf{B}_k = \text{blkdiag}\{\mathbf{B}_k[1], \mathbf{B}_k[2], \ldots, \mathbf{B}_k[N]\}$, and $\mathbf{d}_k = [\mathbf{d}_k[1]^T, \mathbf{d}_k[2]^T, \ldots, \mathbf{d}_k[N]^T]^T$. Using the signal model (2) and the algorithm in [2], the following are evaluated successively in each iteration of the algorithm until convergence:

- MMSE Receiver Filter: $\mathbf{A}_k = \mathbf{B}_k^H\mathbf{H}_k^H(\mathbf{H}_k\mathbf{B}_k\mathbf{B}_k^H\mathbf{H}_k^H + \mathbf{C}_k)^{-1}$

- MSEE-Matrix: $\mathbf{E}_k = (\mathbf{I} + \mathbf{B}_k^H\mathbf{H}_k^H\mathbf{C}_k^{-1}\mathbf{H}_k\mathbf{B}_k)^{-1}$

- Weight Matrix: $\mathbf{W}_k = \mu_k \mathbf{E}_k^{-1}$

- WMMSE Transmit Filter: $\mathbf{\tilde{B}} = (\mathbf{H}_k^H\mathbf{A}_k^H\mathbf{W}_k \mathbf{W}_k + \frac{\text{Tr}(\mathbf{W}_k^H\mathbf{W}_k)}{P_{\text{max}}})^{-1}\mathbf{H}_k^H\mathbf{A}_k^H\mathbf{W}_k$

- Normalized Transmit Filter: $\mathbf{B} = b\mathbf{\tilde{B}}$

Note that because $\mathbf{H}_k$ and $\mathbf{B}_k$ are block-diagonal matrices, the term $\mathbf{H}_k\mathbf{B}_k\mathbf{B}_k^H\mathbf{H}_k^H + \mathbf{C}_k$ in (3) is also block-diagonal. Therefore, computing the MMSE receiver $\mathbf{A}_k$ can be carried out using $N$ matrix-inverse operations on $N_k \times N_k$ matrices, rather than a full-blown inversion of a $NN_k \times NN_k$ matrix. Similarly, computing $\mathbf{E}_k$ and $\mathbf{W}_k$ requires $N$ matrix-inverse operations on $N_k \times N_k$ matrices. The term $\mathbf{H}_k^H\mathbf{A}_k^H\mathbf{W}_k$ in (6) can be written as follows:

$$\mathbf{H}_k^H\mathbf{A}_k^H\mathbf{W}_k = \sum_{k=1}^{K} \mathbf{H}_k^H\mathbf{A}_k^H\mathbf{W}_k \mathbf{A}_k \mathbf{H}_k.$$
The summation in (8) results in a block-diagonal matrix with 
$N$ matrices along the diagonal each of size $M \times M$. Hence, we
 can compute $\hat{\mathbf{B}}$ using $\mathbf{K} \cdot \mathbf{N}$ matrix-inverse operations on $M \times M$
 matrices. After calculating $\hat{\mathbf{B}}$, the beamforming matrices
 must be scaled by the factor $b = \sqrt{\frac{P_{\text{trans}}}{M \cdot \mathbf{B}^H \mathbf{B}}}$ to satisfy the total
 transmit power constraint in each iteration of the algorithm.
The total number of matrix-inverse operations per iteration
 needed to execute WW-OFDMA is:

- $K \cdot \mathbf{N}$ matrix-inverse operations on $N_k \times N_k$ matrices to
  compute $\hat{\mathbf{A}}_k$ using (3);
- $K \cdot \mathbf{N}$ matrix-inverse operations on $N_k \times N_k$ matrices to
  compute $\hat{\mathbf{W}}_k$ using (4), (5); and
- $K \cdot \mathbf{N}$ matrix-inverse operations on $M \times M$ matrices to
  compute $\mathbf{B}$ using (6), (7).
The complexity of WSR maximization in an $N$-subcarrier
 system is therefore only $N$ times that of the single-carrier case.
 We will demonstrate in Section IV that the average number
 of iterations required for this approach to converge is about
 the same as that required by WSRBF-WMMSE. In Approach
 2, we use the expressions (3)-(7) on each subcarrier so that
 the complexity is $N$ times that of the single-carrier case, just
 as in the approach described above. However, performance in
 terms of WSR will be worse than Approach 1, because of the
 sub-optimality of the per-subcarrier design method. We
 therefore propose to optimize jointly over all the subcarriers
 in this work as an extension of WSRBF-WMMSE to MIMO-
 OFDMA systems.

C. Complexity Reduction Using Clustering

Practical MIMO-OFDMA systems employ a large number
 of subcarriers reaching up to 6817 subcarriers in DVB-T.
 We can further reduce the complexity of the algorithm by
 considering the fact that fading on closely spaced subcarriers is
 correlated which makes the beamformers correlated as well.
 Here, we employ clustering [8] where we group subcarriers in
 clusters containing $\nu = N/L$ adjacent subcarriers and design
 a beamformer for the center subcarrier. Clustering has been
 adopted in LTE.

III. STREAM RATE ALLOCATION

A. Problem Description

The algorithm discussed in the last section selects the subset of
 users $\mathcal{K}$ and designs the transmit beamformers $\mathbf{B}_k$ for
 all $k \in \mathcal{K}$ and all subcarriers. The BS transmits multiple
 independent data streams to the $k$-th user on the subcarrier
 $n$, given by the rank of $\mathbf{B}_k[n]$. Often, rank($\mathbf{B}_k[n]$) = $N_k$ so
 that the total rate for user $k$ on the $n$-th subcarrier, $\check{R}_{k,n}$, is
 split among $N_k$ streams, or $\check{R}_{k,n} = \sum_{l=1}^{N_k} \check{R}_{k,l}[n]$ where
 $\check{R}_{k,l}[n]$ is the rate allocated to stream $l$ of user $k$ on subcarrier
 $n$. In theory, there are an infinite number of rate tuples
 $\check{R}_k = [\check{R}_{k,1}, \ldots, \check{R}_{k,N_k}]$ with the same total rate $\check{R}_k$, as
 explained below. In fact, we will view the channel between
 the multiple data streams transmitted from the BS to user
 $k$ on subcarrier $n$ and receiver $k$ as a multi-access channel
 (MAC), having a corresponding achievable rate region that
 encompasses all feasible rate tuples $\mathbf{R}_k$. In practice, only
 a finite number of adaptive modulation and coding (AMC)
 modes are available for each data stream, and this results in
 a constellation of available rate tuples $\mathbf{R}_k$. After adjusting for
 the SNR gap to capacity, the AMC modes that are feasible
 are determined by the corresponding $\mathbf{R}_k$ tuples that lie within
 the theoretically achievable rate region. We then pick the one
 which gives the largest total rate $R_k = \sum_{l=1}^{N_k} R_{k,l}$. If there
 is more than one such point, we will choose the one that is
 furthest from the achievable rate region boundary.

B. Multi-Access Channel Analysis

We define $\mathbf{d}_k = [d_{k,1}, \ldots, d_{k,N_k}]^T$ as the vector of symbols
 transmitted to user $k$ (on the $n$-th subcarrier) and $\mathbf{G}_k =
 [\mathbf{g}_{k,1}, \ldots, \mathbf{g}_{k,N_k}] = \mathbf{H}_k \mathbf{B}_k$. Then, (2) can be re-written as

$$y_k = \sum_{l=1}^{N_k} \mathbf{g}_{k,l} \mathbf{d}_{k,l} + \sum_{j=1,j\neq k}^{K} \mathbf{H}_{k,j} \mathbf{d}_j + \mathbf{w}_k.$$  \hspace{1cm} (9)

By considering the multi-user interference (MUI) term
 $\sum_{j\neq k} \mathbf{H}_{k,j} \mathbf{d}_j$ as part of the Gaussian noise, we can further
 write

$$y_k = \sum_{l=1}^{N_k} \mathbf{g}_{k,l} \mathbf{d}_{k,l} + \mathbf{v}_k = \mathbf{G}_k \mathbf{d}_k + \mathbf{v}_k,$$  \hspace{1cm} (10)

where $\mathbf{v}_k \sim \mathcal{N}(0, \mathbf{C}_k)$ is coloured and Gaussian, assuming
 ideal Gaussian coding for all users. Expression (10) is identical
to that of a vector Gaussian MAC, in which user $l$ transmits the
 symbol $\mathbf{d}_{k,l}$ over a vector channel $\mathbf{g}_{k,l}$. This resemblance of
 the downlink channel between the base station and a receiver
 with multiple antennas to a Gaussian MAC is a key realization
 used in the rate allocation algorithm to follow. Given the
 equivalence between (10) and the vector Gaussian MAC with
 $N_k$ users, all rate tuples $(\check{R}_{k,1}, \ldots, \check{R}_{k,N_k})$ in the interior of
 the polymatroid MAC capacity region defined by the channels
 $(\mathbf{g}_{k,1}, \ldots, \mathbf{g}_{k,N_k})$ and the distribution of $\mathbf{v}_k$ are achievable. All
 points on the maximum sum-rate surface yield the designed
 total rate $\check{R}_k$ for user $k$, and in theory any of these rate tuples
 can be used with the beamforming matrix $\mathbf{B}_k$ designed by
 the proposed algorithm. The maximum sum-rate surface is
 defined by $N_k!$ vertices, each one corresponding to one order
 of successive interference cancellation (SIC). For the decoding
 order $1, \ldots, N_k$, the $l$-th entry of the rate tuple is

$$\check{R}_{k,l} = \log_2 (1 + \mathbf{g}_{k,l}^H (\mathbf{C}^{SIC}_{k,l})^{-1} \mathbf{g}_{k,l}).$$  \hspace{1cm} (11)

where $\mathbf{C}_{k,l}^{SIC}$ is the covariance matrix of the interference seen
 by $\mathbf{d}_{k,l}$, which includes AWGN, inter-user interference, and
 intra-user (inter-stream) interference. Therefore we have

$$\mathbf{C}_{k,l}^{SIC} = \mathbf{I} + \sum_{j=1,j\neq k}^{K} \mathbf{H}_{k,j} \mathbf{B}_j^H \mathbf{H}_k^H + \sum_{i=l+1}^{N_k} \mathbf{g}_{k,i} \mathbf{g}_{k,i}^H.$$  \hspace{1cm} (12)


\footnote{The subcarrier index $n$ is dropped in the rest of this section as it is irrelevant.}
In theory, any point on the maximum sum rate surface can be used with optimal decoding, and any vertex with SIC having the right decoding order. In practice, there are two problems that need to be addressed: (1) The rates $R_{k,l}$ are not achievable using practical modulation and coding — there is a gap to capacity which is not negligible; (2) as pointed out earlier, a system cannot transmit at arbitrary rates. Instead only a finite number of rate values are available. The algorithm described next seeks to address both problems.

### C. Rate Allocation Method

Depending on the structure of the receiver, the solution to the rate-splitting problem may differ. In this work, we have considered two receiver structures: ML and successive interference cancellation (SIC).

#### 1) ML Receiver

Denote the SNR gap to capacity in the $l$-th stream of the $k$-th user by $\Gamma_{k,l}$ [9], [10], and let the SINR of that stream be $\gamma_{k,l}$. Then, for high values of $\gamma_{k,l}$, the practically achievable rate can be approximated as [10]:

$$R_{k,l} \approx \log_2(1 + \gamma_{k,l})$$

and hence

$$\tilde{R}_{k,l} \approx R_{k,l} + \log_2(\Gamma_{k,l}).$$

Equation (14) tells us that if we want to transmit at the rate $R_{k,l}$, we need to design a system that has a theoretical achievable rate about $\log_2(\Gamma_{k,l})$ bits higher. Problem 2 above is addressed in conjunction with Problem 1 by using (14) to derive the $\tilde{R}_{k,l}$ corresponding to each allowable practical rate $R_{k,l} = r \cdot \log_2 M$, where $r$ is the code rate and $M$-ary modulation is used. Assume the transmitter can choose among $P$ $M$-ary square QAM constellations with $M \in \{2^{b_1}, 2^{b_2}, \ldots, 2^{b_P}\}$ and $Q$ coding schemes $\{d_1, d_2, \ldots, d_Q\}$ each associated with a code rate $r(d_q) \in (0, 1]$ and a coding gain $G_c(d_q)$. The SNR gap due to the use of $2^{b_p}$-ary QAM inputs rather than Gaussian inputs is given by

$$\Gamma_{k,l} = \frac{1}{3} \left( Q^{-1} \left( \frac{b_p P_b}{(1 - 2^{-b_p/2})} \right) \right)^2,$$

[9], where $P_b$ is the bit error rate (BER). If the coding gain provided by a coding scheme $d_q$ at a certain desired error probability for stream $l$ of user $k$ is denoted $G_{c,k}(d_q)$, then we can write the gap to capacity for each feasible choice of coding and modulation formats as

$$\Gamma_{k,l} = \frac{\Gamma_{k,l}^{\text{mod}}}{G_{c,k}(d_q)}.$$  

The literature contains tables of coding gains achieved by different coding schemes at a given error probability. Having defined the gap to capacity, the constellation $\mathcal{V}_k$ of allowable theoretical rate vectors $(\tilde{R}_{k,1}, \tilde{R}_{k,2}, \ldots, \tilde{R}_{k,N_k})$ can be found from the corresponding set of allowable practical rate vectors $(R_{k,1}, R_{k,2}, \ldots, R_{k,N_k})$. By finding the members of $\mathcal{V}_k$ contained within the MAC capacity region with channels $g_{k,1}, g_{k,2}, \ldots, g_{k,N_k}$, denoted as $C_{MAC}^M(H_k, B)$ to reveal the dependence of the capacity region on the $k$-th downlink channel and the set of designed precoders $B = \{B_1, B_2, \ldots, B_K\}$, we obtain the subset of allowable rate vectors. From this subset of $\mathcal{V}_k$, we choose the operating point that yields the highest practical sum rate $\sum_l \tilde{R}_{k,l}$. If there are more than one operating point giving this rate, then we choose the one that is furthest from the capacity boundary, in order to provide the largest margin of error.

To illustrate the above concepts, consider Fig. 1 which simulates the rate region corresponding to a given channel realization of the first user in a system with $M = 4$, $K = 2$, and $N_k = 2$ employing uncoded modulation at SNR $= 20$ dB and $P_b = 10^{-3}$. We assume the BS has access to four modulation schemes $\{4, 16, 64, 256\}$-QAM. The points shown are the possible operating points with the given modulation schemes. The coordinates of the operating points were calculated using (14). Since uncoded modulation schemes are used, we have $\Gamma_{k,l} = \Gamma_{k,l}^{\text{mod}}$ which was calculated to be around $2^{1.6}$ for the used modulation schemes. The set $X_1 \cup X_2$ contains all rate maximizing points in $\mathcal{V}_1$. The point in the set $X_1$ corresponds to $R_{1,1} = 2$ and $R_{1,2} = 4$, while the point in $X_2$ corresponds to $R_{1,1} = 4$ and $R_{1,2} = 2$. The maximum practical rate available to this user is therefore 6 bits. In this example, the algorithm would choose the point belonging to $X_1$ as it is the furthest from the boundaries of the rate region which would allow for the largest error margin.

#### 2) SIC Receiver

A SIC receiver allows for different possible decoding orders. Therefore, for each practical rate vector, there exists $N_k!$ theoretical rate vectors corresponding to all the possible decoding orders. In this case, the covariance matrix of the interference seen by $d_{k,l}$ is given by (12) assuming a SIC receiver acting on streams 1 through $N_k$ successively. We note that the last term on the RHS of (12) depends on decoding order. By introducing another SNR gap parameter $\Gamma_{k,l}^{\text{dec}}$, the practically achievable rate becomes

$$R_{k,l} = \log_2 \left( 1 + \frac{1}{\Gamma_{k,l}^{\text{dec}}(C_{E,N_k}^{SI})^{-1} g_{k,l}} \right),$$  

where $C_{E,N_k}^{SI}$ is the covariance matrix of the interference seen by $d_{k,l}$.
where now
\[ \Gamma_{k,l} = \frac{\Gamma_{k,l}^{\text{mod}} \Gamma_{k,l}^{\text{dec}}}{G_{k,l}^H(d_q)} \].

(17)

The new parameter \( \Gamma_{k,l}^{\text{dec}} \) represents the increase in SNR required to transmit rate \( \tilde{R}_{k,l} \) as the \( l \)-th decoded stream, rather than the last decoded stream. By comparing (16) and
\[ \tilde{R}_{k,l} = \log_2 \left( 1 + \frac{G_{k,l}^H(d_q) H_{k,l} \left(C_{k,l}^{\text{SIC}}\right)^{-1} G_{k,l}}{\Gamma_{k,l}^{\text{mod}} \Gamma_{k,l}^{\text{dec}} G_{k,l}^H(d_q)} \right), \]

we conclude that
\[ \Gamma_{k,l}^{\text{dec}} = \frac{H_{k,l} \left(C_{k,l}^{\text{SIC}}\right)^{-1} G_{k,l}}{G_{k,l}^H \left(C_{k,l}^{\text{SIC}}\right)^{-1} G_{k,l}} \]

(19)
is the SNR gap at stream \( l \) due to its position in the decoding order. One should note that if a SIC receiver is used, points such as those in Fig. 2 are not achievable without time-sharing.

---

**IV. SIMULATION RESULTS**

In this section we perform different experiments to evaluate the performance of WW-OFDMA as well as the proposed rate allocation algorithm. In all the following, we assume the BS has access to four modulation schemes \{4, 16, 64, 256\}-QAM, \( M = 4 \), and \( \mu_k = 1 \), unless mentioned otherwise. For WW-OFDMA, we chose the transmit matched filter as an initialization for the beamformers, i.e., \( B_k = \alpha H_k^H \) where \( \alpha \) is selected to satisfy the transmit power constraint. The channel gains are assumed to be ZMCSCG with variance \( \sigma_k^2 \). Quasi-static fading is assumed where the channels are constant over the duration of a codeword and change independently between codewords. The curves are generated by averaging over 1000 channel realizations. SNR is equal to \( \sigma_k^2 \) since both the signal and noise powers are normalized to have unit power.

**A. WW-OFDMA**

Fig. 3 shows the number of iterations required for the convergence of WW-OFDMA for different parameters. It can be seen that the algorithm needs about 5 iterations to converge, similar to WSRBF-WMMSE, without the need to solve a geometric program as does [3], [4], [5], [6]. Fig. 4 compares the achieved throughput per subcarrier (WSR \( /N \)) by DPC and WW-OFDMA for \( K \in \{4, 20\} \) and \( N_k = 1 \). The DPC bound was obtained by applying the algorithm in [11] to parallel channels. We allowed WW-OFDMA to run for 10 iterations only which are enough for convergence, as shown in Fig. 3. We observe that WW-OFDMA exhibits near-optimal performance. In Fig. 4(a), it is seen that WW-OFDMA sees a loss starting at \( \text{SNR} = 15 \text{ dB} \) compared to DPC. This loss is less severe in Fig. 4(b) due to multi-user diversity. It is worth mentioning that a different initializing filter might lead to an improved performance. This is because the WSR maximization problem is nonconvex and the initial filter used determines whether WW-OFDMA will converge to a local or global solution.

---

**Fig. 2.** An example with multiple solutions. \( M = 4, K = 2, k = 1, N_k = 2, \text{SNR} = 20 \text{ dB}, \text{and } P_0 = 10^{-3} \).

**Fig. 3.** Number of iterations needed for WW-OFDMA to converge for a fixed channel realization under different parameters. \( M = 4, L = 10, \text{SNR} = 20 \text{ dB} \).

**Fig. 4.** Comparing the average throughput achieved by WW-OFDMA and DPC. \( M = 4, K \in \{4, 20\}, N_k = 1, N = 32, \text{and } L = 10 \). (a) \( K = 4 \) (b) \( K = 20 \).
size of the clusters $\nu$ increases, more subcarriers use the same beamformer which leads to increased interference among users and decreased throughput.

B. Stream Rate Allocation

The simulations in this section are for a given subcarrier which is exposed to flat fading as opposed to frequency-selective fading due to OFDM.

Fig. 6 shows the average WSR achieved by WW-OFDMA and our rate allocation algorithm for $M = 4$, $K = 20$, and $P_0 = 10^{-3}$. For the stream rate allocation algorithm, we assume that two coding schemes $d_1$ and $d_2$ are available with $r(d_1) = r(d_2) = 1/2$ and $G''(d_1) = 4$ dB, $G''(d_2) = 5.5$ dB. We estimate the rate for systems employing ML and SIC receivers. We observe that our algorithm traces the ideal rate of WW-OFDMA for both type of receivers. In the ML receiver case, the throughput of the proposed algorithm caps at around 32 bits for SNR $\geq 15$ dB. This can be explained by noting that at high SNR with $M < K$, WW-OFDMA selects $M$ users on each subcarrier, that are nearly orthogonal and have high channel gains. Also, the largest modulation scheme available is 256-QAM, and each user can receive a maximum of $N_k = 2$ symbols. Each stream carries a maximum of 4 information bits due to the rate of the employed coding schemes. Hence, the maximum throughput possible at high SNR is 32 bits. For the SIC case, the receiver is not able to achieve all the points in the rate region, and the gap to capacity due to the decoding order reduces the number of points in the region which limits the number of total received streams that are assigned practical rates to about 4 data streams, and hence the sum-rate reaches a maximum of 16 bits.

V. CONCLUSION

This paper proposes a practical WSR maximization algorithm using linear processing for MIMO-OFDMA systems. The increase of complexity induced by WW-OFDMA relative to its single-carrier counterpart was shown to scale only linearly with the number of subcarriers offering an attractive solution for the WSR maximization problem for multi-carrier systems. We also proposed a rate allocation algorithm to translate theoretical data rates achieved by the beamformers designed using a WSR maximizing algorithm into practical AMC modes for each user, for a given desired bit error probability. The SNR gap to capacity concept was extended to include the effect of order of detection in SIC and coding gain of the employed coding scheme; it was then used to map available AMC modes to the space of allowed theoretical rates.

REFERENCES