Practical Bit Loading Schemes for Multi-Antenna Multi-User Wireless OFDM Systems

Diego Bartolomé and Ana I. Pérez-Neira

I. INTRODUCTION

Currently, wireless multi-carrier communications are widely deployed, and multiple antennas further enhance the system performance. With multiple users, the problem of allocating the users into subcarriers, antennas, and performing the convenient integer bit allocation is NP-complete [1]. We focus on the Access Point (AP) in the downlink of an OFDM-based Wireless LAN, such as IEEE 802.11a. Due to the NP-completeness, the problem can be separated into two parts, namely the user grouping and the beamforming, power, and bit allocation for each of the groups at all the subcarriers.

In a realistic scenario, the AP shall distribute the users into groups of Q for each subcarrier, note that K > Q. Since this problem is NP-complete [1], suboptimum solutions are adequate, see e.g. [2] for an example of the uplink of an SDMA/TDMA system. In [3] the authors extend [2] to take into account several QoS parameters. In this paper, we justify and propose a greedy algorithm based on the normalized scalar product to allocate the users into groups.

Then, for each group, it is meaningful to separate the transmit beamforming and the power (and bit) allocation [4]. A well-suited beamforming criterion is Zero Forcing (ZF) [5], which provides a reasonable performance loss with respect to optimum downlink beamforming e.g. [6] or dirty paper precoding [7]. On top of it, the AP shall perform the power (and bit) allocation. Without multiple antennas, several bit allocation strategies for multi-carrier systems have been developed since [8], see e.g. [9], [10], [11], or [12] and references therein. The use of a ZF transmit beamforming forces their modification to take into account the special characteristics of the spatial diversity system. Therefore, we develop in this paper extensions of well-known algorithms for this multi-antenna multi-user multi-carrier system. Furthermore, we describe simple power reuse schemes, and propose a solution to reduce the amount of signaling, which affects the fairness.

II. PROBLEM STATEMENT

Boldface capital (lowercase) letters refer to matrices (vectors). The conjugate transpose of a vector a is aH and the element at row i and column j of a matrix A is denoted by [A]i,j (A). The square matrix with the diagonal given by a1, a2, ..., an is diag(a1, a2, ..., an). The cardinality of the set K is expressed as |K|. det(A) is the determinant of A, and tr(A) is the trace. The vector 1K has zeros at all positions but the kth.

The system is summarized in Fig. 1. The ultimate purpose of the scheduling is to distribute the K users in the cell into groups at every subcarrier, so that they can be served simultaneously by the Q-antenna AP (K > Q). Since multi-carrier modulations are well-known, the signal model in this section is devoted exclusively to the frequency domain representation. Moreover, since the optimization procedures are performed instantaneously, we omit the time index in (1). Assuming that the N subcarriers have their particular set of users Kn to be served, the signal at subcarrier n is given by

\[ y(K_n) = H(K_n)B(K_n)s(K_n) + w(K_n) \in C^{Q \times 1}, \]

where Kn emphasizes that the signal model is expressed for subcarrier n and for the simultaneously-served users in Kn. The kth position of y(K_n) (s(K_n)) is the received (transmitted) signal for user k in the set Kn at the subcarrier n. The transmit beamvectors are gathered in B(K_n) = [b1(K_n), b2(K_n), ..., bK(K_n)] ∈ CQ×K. H(K_n) is the |K_n| × Q complex flat-fading channel matrix at the nth subcarrier, the kth row of which contains the 1 × Q vector of the channel gains for the kth user at the nth subcarrier, i.e. h_{k,n}^T, which is obtained evaluating the Fourier transform of the L-tap channel vector h_k at the nth subcarrier, i.e.

\[ f_n^H h_k = \begin{bmatrix} 1 & \exp(-j2\pi n) & \ldots & \exp(-j2\pi n(L-1)) \end{bmatrix}. \]

We assume that the channels h_k are independent, and perfectly known at the AP. The noise w(K_n) is independent zero-mean complex Gaussian random variables with variance σ^2.
As in other papers such as [5] or [13], we assume a ZF transmit beamforming, which is equivalent to the MMSE for a low number of users and also for high SNR [14]. Moreover, it provides a reasonable degradation with respect to the optimum sum capacity as shown in [15]. ZF implies the beamforming matrix also contains the power allocation for the union of the possible constellations together with the power by substituting (4) into (7), i.e.

\[
\beta_{k,n}^2 = \frac{\sigma_n^2 \gamma_{k,n}}{c_2 \alpha_k^2(K_n)} \log \left( \frac{c_1}{BER_t} \right),
\]

which can be inserted into (6) to compute the total used power. The objective in (5)-(8) involves mainly two tasks: i) the grouping of the users in the sets \( K_n \) at the N subcarriers, and ii) the space-frequency bit allocation. In the remainder of this section, we justify the use of the scalar product to allocate the users at the subcarriers.

A. Towards a simple tool to the user clustering

As stated, the objective in (5)-(8) is NP-complete [1], which means that it cannot be solved in polynomial time, i.e. the complexity increases exponentially with the number of variables. It is shown in [2] by graph theory that the problem of minimizing the length of an SDMA/TDMA frame, while ensuring a minimum SINR for each terminal, is NP-complete. In our case, the proof follows directly from linear programming, in which several combinatorial problems are known to be NP-complete, among others the Knapsack. Without going into the details, the maximization in (5)-(8) can be mapped into the well-known Knapsack problem, thus it is NP-complete.

To build real-time channel-adapted realizable schedulers, suboptimum algorithms are therefore motivated. Several papers have proposed the scalar product of the channel vectors among the users as a well-suited approach for the design of schedulers, see e.g. [4]. We prove next that this approach is a valid path to separate the users intelligently. A similar concept of user separability is developed e.g. in [17], where heuristic algorithms are developed to study the impact of smart antennas on the Medium Access Control (MAC) layer, with especial emphasis on the limited availability of resources, i.e. they assume that only a finite set of beams is available. A related concept based exclusively on power capture threshold is [18] for a sectorized AP. However, to the best of our knowledge, the normalized scalar product has not been yet shown to be a suited technique to separate the users.

To show the previous statement, we concentrate on a single subcarrier and omit the index \( K_n \) for simplicity. \( K = 2 \) users are considered, denoted by the index \( k \in \{1, 2\} \). Their channels are expressed as \( h_k \), and are gathered in the matrix \( H \). If the two users use the same signal mapping, it follows from (9) that \( \sum \beta_k^2 \propto \sum 1/\alpha_k^2 = tr(\text{HH}^H) \). For the proof, we shall first compute the determinant of the matrix.
HH', where $H = [h_1, h_2]^T$. After some manipulations, it yields $\text{det}(HH') = ||h_1||^2||h_2||^2 - ||h_1^H h_2||^2$. As stated, the consumed power is determined by the trace of the matrix $(HH')^{-1}$, so we find

$$tr((HH')^{-1}) = \frac{||h_1||^2 + ||h_2||^2}{\text{det}(HH')} = \frac{||h_1||^2 - ||h_2||^2}{1 - c_{1,2}},$$

where $c_{1,2} = \frac{||h_1^H h_2||}{||h_1|| ||h_2||}$, and the second equality is obtained by dividing both the numerator and the denominator by $(||h_1|| ||h_2||)^2$. If there were $Q = 2$ antennas, $c_{1,2}$ would reflect the cosine of the angle between $h_1$ and $h_2$. Anyway, $c_{1,2}$ is the square of the normalized scalar product between $h_1$ and $h_2$ defined accordingly, and its range is $0 \leq c_{1,2} \leq 1$. The lower bound occurs if $h_1$ and $h_2$ are orthogonal, and the upper bound when $h_1 = h_2$. With all this,

$$\lim_{||h_1|| \to \infty} tr((HH')^{-1}) = \frac{||h_2||^2}{1 - c_{1,2}}$$

and it is easy to see that it is bounded if $h_1 \neq h_2$. The same situation occurs if the limit is calculated for $||h_1|| \to \infty$. However, if we compute the limit when $h_1 \to h_2$, or equivalently when $c_{1,2} \to 1$, it yields

$$\lim_{c_{1,2} \to 1} tr((HH')^{-1}) = \infty.$$ 

Therefore, it is more critical to separate those users coming from the same zone of space rather than using the norm of their channel vector as a performance measure to allocate users. This justifies the use of a measure like $c_{1,2}$ as a way to allocate the users into the subcarriers. If more users form the matrix $H$, the cost is determined by the one with a highest $c_{i,j}$.

### III. SPACE-FREQUENCY MULTI-USER SCHEDULING

With the previous results, we propose a suboptimum yet very simple real-time approach based on the normalized scalar product to allocate the users into subcarriers in groups of $Q$, and then we deal with the spatial power and bit allocation, which is an extension to OFDM of the Maximum Sum Rate spatial bit allocation for single carrier developed in [19]. Finally, we propose practical schemes for comparison.

#### A. User-subcarrier assignment based on the scalar product

The initialization is the following: the AP computes the cost for each pair of users, $i$ and $j$, for all the subcarriers, thus $\mathcal{K}_n = \{i, j\} \forall n$. However, since the channel is generally frequency-selective, the separation of the users depends on the particular subcarrier. Therefore, the associated cost of putting user $i$ and user $j$ together is determined by the maximum cost among the subcarriers. Indeed, this is the worst case among user $i$ and $j$, which is the limiting factor. Then,

$$c_{i,j} = \max_n \frac{||h_i^H h_j||}{||h_i|| ||h_j||}, \forall i \neq j$$

are computed for all pairs of users. The AP sorts then these values in descending order, and starts assigning the users with higher costs at adjacent subcarriers, until no carriers are left. After this procedure, each carrier is filled with a single user, and the AP shall fill the subcarriers until there are $Q$ users pre-allocated per subcarrier. One could use rather cost-extensive approaches such as those developed in [5], but the scalar product is a good and very simple option. The procedure is summarized in Table I, and it is very intuitive. When the AP has $Q$ pre-allocated users per subcarrier, it applies the MMSR.

#### B. Multi-antenna Multi-carrier Maximum Sum Rate (MMSR)

The space-frequency bit allocation has some differences with respect to traditional multi-carrier bit loading, e.g. the channels change accordingly when the users that are simultaneously served change, see also [19]. In realistic scenarios with several users, if the problem is not feasible, then the AP has to perform also the admission control, i.e. choose the users that will be served. Essentially, two strategies can be found in the single antenna bit loading literature in Section I, namely bit filling and bit removal. The former adds a bit to the user/subcarrier providing the lowest increase in total power, and bit removal schemes remove the most penalizing bit until the power constraint is fulfilled. With multiple antennas, it is not possible to do strict bit filling algorithms, since the interactions among the users that are being simultaneously served are crucial. For instance, the user with best channel might not even have a good channel when grouped together in an SDMA scheme [19].

The number of bits for user $k$ at subcarrier $n$ is $m_k(K_n)$, except for the $l$th, which changes the number to $m_l(K_n)$ instead of $m_l(K_n)$, where $m_l(K_n) > m_l(K_n)$. The power saving $p_{l,n}(m_l(K_n), m_l(K_n))$ can be approximated as $\sigma^2_l 2^{m_l(K_n)} - 2^{m_l(K_n)}/\alpha^2_l(K_n)$ if $m_l(K_n) \in \mathcal{M}$, and as $\sum_{k \in \mathcal{K}_n} \sigma^2_k n_k a^{m_n(K_n)}/\alpha^2_k(K_n) - \sum_{k \in \mathcal{K}_n} \sigma^2_k n_k a^{m_n(K_n)}/\alpha^2_k(K_n)$ if $m_l(K_n) \notin \mathcal{M}$. $\mathcal{K}_n$ gathers all the users but the $l$th and the equivalent channels $\tilde{\alpha}_k$ refer to the users in $\mathcal{K}_n$. In fact, this is not the exact saving that is obtained because if a user is removed, then the rest have the chance to increase their modulation index. However, the complexity reduction of this approximation justifies its use.

The MMSR in Table II is based on a bit removal technique, but it is aided by a bit filling scheme, performed when the spatial channel gains $\alpha_k$ change, i.e. whenever the set of active users varies. Briefly, the MMSR first tries to serve all the users in the set $\mathcal{K}_n$, $\forall n$ obtained by the user-subcarrier clustering method with the highest modulation in $\mathcal{M}$ at steps 2-4. If the power constraint in (6) is not fulfilled (step 5), the scheduler decides which user among all the carriers should reduce the constellation size or which user should not be served. Since the number of bits shall be reduced, the scheduler selects the user having a maximum incremental cost of using a lower modulation, i.e. the user that saves more power if the bit rate is

<p>| TABLE I |</p>
<table>
<thead>
<tr>
<th>USER CLUSTERING BASED ON SCALAR PRODUCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Set $n = 0$. The users for subcarrier $n$ are collected in $\mathcal{K}_n$, which has been initialized.</td>
</tr>
<tr>
<td>2. If $n = N - 1$, finish. Otherwise, do $n \rightarrow n + 1$.</td>
</tr>
<tr>
<td>3. Select $k : \min_{x \in \mathcal{K}_n} \epsilon_k(K_n)$. Add user $k$ to $\mathcal{K}_n$, i.e. $\mathcal{K}_n \leftarrow \mathcal{K}_n + k$.</td>
</tr>
<tr>
<td>4. If $</td>
</tr>
</tbody>
</table>
The complexity of (5)-(8) might be reduced by separating the problem into smaller ones, i.e. solve the problem independently at each subcarrier. With a low $K$, we could even perform the Exhaustive Search (ES) for each subcarrier. Since a significant amount of power might remain unused, some kind of power reuse scheme is needed. The procedure is very simple: we perform the ES at each subcarrier with an available power of $P_T/N + P_n$, where $P_n$ gathers the accumulated unused power for the subcarriers previous to the $n$th. Then, an extremely simple scheme is obtained, when compared to the iterative (and far more complex) power redistribution routine for the eigenmodes of single link MIMO system in [20].

The same idea is valid for the previous MMSR performed independently for each subcarrier. The simplest scheme is the Random Approach, which selects a random combination of the set of active users and then performs with them the proposed MMSR spatial bit allocation for each subcarrier with the power reuse. We shall also consider Opportunistic communications. With multiple antennas, we consider that it refers to the fact that the spatial diversity is used to enhance the SNR at the receiver, but there is no (user) multiplexing gain. Only the user with the best norm of the channel, that with $\max_k ||h_{ik}||$, is scheduled for transmission, with the highest number of bits per symbol satisfying the BER constraint.

**D. Reducing the signaling needs**

The amount of overhead required might have relevance especially when the number of transmitted OFDM symbols is low, and might deeply penalize the performance. If $b$ bits are required to transmit the desired constellation to the users, a total number of $bT = Q \times N \times b$ bits are needed for signaling at every burst. Among other simple practical options to reduce the signaling needs, the AP might force an equal signal mapping for all the users at the same subcarrier, i.e. the Multi-carrier Maximization of the Minimum Rate (MMMR). This might reduce the overhead by a maximum factor of $Q$ (upper bound). In fact, by using the same mapping for all the users we guarantee that the users being served receive the same rate. This ensures the fairness among users if they are homogeneous (or pay the same price for the service). In some sense, we maximize the number of users that are served but the global performance might be penalized. Fairness issues have been rarely evaluated in the literature, see e.g. [21], or the simplified approach in [14]. A theoretical study of fairness in multi-antenna multi-user channels is conducted in [22]. The MMMR alternative can be expressed as a max-min problem, i.e.

$$\max_{m_k(k_n)} \min_{0 \leq n \leq N-1} \sum_{k} m_k(k_n), \quad \text{s.t.} \quad (6), (7), \text{and (8)}.$$  

Again, the optimum solution for this problem implies the exhaustive search among all the users and all the number of bits. The Multi-carrier Maximization of the Minimum Rate (MMMR) algorithm is summarized in Table II, which consists essentially of the same steps as the MMSR, but in this case, all the users at subcarrier $n$ lower their modulation size when the selected user is at the same subcarrier, see step 8.

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>SPACE-FREQUENCY BIT ALLOCATION: MMSR AND MMMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The set $k_n$ is obtained by the user-subcarrier assignment.</td>
<td></td>
</tr>
<tr>
<td>2. Set $m_n(k_n) = \max M_k, \forall k \in k_n, 0 \leq n \leq N-1$.</td>
<td></td>
</tr>
<tr>
<td>3. Build $H(k_n)$ and compute $\sigma_{k_n}^2$.</td>
<td></td>
</tr>
<tr>
<td>4. Compute $\beta_n^2(k_n)$ according to (9), and the total used power $P_T = \sum_{n=0}^{N-1} \sum_{k \in k_n} \beta_n^2(k_n)$.</td>
<td></td>
</tr>
<tr>
<td>5. If $P_T \leq P_T$, or $k_n = 0$, then finish.</td>
<td></td>
</tr>
<tr>
<td>6. Compute $p_{n,k} = m_n^2(k_n, m_k^2(k_n)) \cdot 0 \leq n \leq N-1, k \in k_n$, where $m_n^2(k_n)$ is the current mapping and $m_n^2(k_n)$ the lower one in $M$.</td>
<td></td>
</tr>
<tr>
<td>7. Select ${n,k} : \max_n \max_k m_{n,k} \cdot m_k^2(k_n, m_k^2(k_n))$, and reduce the number of bits $m_n^2(k_n) - m_k^2(k_n)$.</td>
<td></td>
</tr>
<tr>
<td>8. Only for MMMR: $m_n^2(k_n) - m_n^2(k_n) \forall k \in k_n$.</td>
<td></td>
</tr>
<tr>
<td>9. If $m_n^2(k_n) \in M$, go to step 4. Else, $k_n - k$, set $m_n^2(k_n) = \max M, \forall k \in k_n$, and go to step 3.</td>
<td></td>
</tr>
</tbody>
</table>

### IV. PERFORMANCE EVALUATION

Simulation are conducted for a typical office environment with 50 ns average rms delay spread for OFDM-based Wireless LAN, model A in [23]. We assume for simulation purposes that the noise is equal for all subcarriers $\sigma_{k,n}^2 = \sigma^2, \forall n, k$. We define the SNR as the ratio $P_T / \sigma^2$. We assume $N = 64$ subcarriers, and QAM constellations with $M = \{2, 4, 6\}$ bits per symbol. The BER target is $1e-4$.

First, we compare the MMSR to the strategies with a simple power reuse. We plot in Fig. 2 the average throughput at the physical layer in terms of number of bits per symbol per subcarrier vs. the SNR, when $K = 5$ and $Q = 3$. We see that the Exhaustive Search (ES) scheme, with or without power reuse, outperforms the other methods in the high signal to noise ratio range at the expense of a prohibitive computational complexity when $K$ is high. The performance of the MMSR has practically no degradation at low SNR, but differences are higher when the SNR increases. The simple scheme with power reuse is very close to the globally computed maximum sum rate. If one performed the schemes individually per subcarrier and without power reuse, some power would be wasted, thus the rate would be penalized. The suboptimum user-subcarrier clustering based on the scalar product provides a reasonable trade-off between performance and complexity.

Second, we simulate a more realistic scenario with $K = 20$ users and the same number of antennas. We plot in Fig. 3 the average throughput at the physical layer in terms of number of bits per symbol per subcarrier vs. the SNR. We see again that the performance with the globally computed algorithms is very close to the use of the simple power reuse. If we impose equal constellation for all users at a certain subcarrier, i.e. the MMMR, the performance is penalized with respect to the...
MMMR, but the amount of signaling is also reduced and the AP guarantees fairness. In any case, a noticeable gain is achieved with respect to a random selection of users. Finally, one can see that opportunistic communications yields a low throughput at high SNR because of the limitation of the scheduling of a single user. However, at low SNR where the noise is dominates the performance tends to be optimum, see [7] or [24]. In that region, typically a single user is scheduled per subcarrier at most, therefore the MMMR is equivalent to the MMSR.

V. CONCLUSIONS

We have studied the practical implementation of bit allocation techniques with a combination of space and frequency diversity. We have reviewed the particular issues that arise with the spatial dimension, and have shown that the objective is NP-complete. After proving that solutions based on the scalar product are good candidates for a fast and realizable cross-layer scheduler, we have described a very simple strategy. We have proposed a mechanism to reduce the huge signaling needs of the multi-user multi-carrier spatial bit allocation, which has also fairness implications. By means of realistic simulations for typical indoor Wireless LAN environments, we have shown that simple practical schemes might be adequate in the design of schedulers. Besides the trade-off between performance and complexity, there exits the trade-off between performance and signaling, and between global and the individual needs. For this reasons, the final choice strongly depends on the AP.

REFERENCES


Fig. 2. Comparison in terms of throughput vs. the ratio $SNR = P_t/\sigma^2$ to evaluate the power reuse, with $K = 5$ and $Q = 3$.

Fig. 3. Comparison in terms of throughput vs. the ratio $SNR = P_t/\sigma^2$ in a more realistic case, $K = 20$ and $Q = 3$.