Round-Robin Scheduling for Orthogonal Beamforming with Limited Feedback

Claude Simon, *Member, IEEE*, and Geert Leus, *Senior Member, IEEE*

***Abstract*—We propose a round-robin scheduling algorithm for orthogonal beamforming with a strict signal to interference-plus-noise ratio (SINR) constraint and limited feedback. The presented algorithm aims at scheduling the users at identical slots over different blocks, in order to reduce the necessary scheduling overhead, and to minimize the maximum delay between serving the same user. Thus, the presented algorithm is especially suited for real-time multimedia traffic. The algorithm allocates the users using orthogonal beamforming based on the quantized feedback provided by the users. The quantized feedback consists of the estimated power that is necessary to fulfill a predefined SINR constraint. Further, we propose an algorithm to design codebooks to quantize the estimated power. Using the feedback, the base station redistributes power from users with spare power to users that lack power so that they fulfill their SINR constraints. The performance of the algorithm is demonstrated through simulations.**

***Index Terms*—Scheduling, limited feedback, beamforming.**

I. Introduction

Some of the key drivers of wireless communication are delay-critical services like audio and video communication. Of special interest for these services is the vector Gaussian broadcast channel where the base station incorporates multiple antennas but the individual users just have a single antenna. The vector Gaussian broadcast channel promises large sum rates [1] if perfect channel state information (CSI) is available at the base station. CSI is easily acquired at the user side through training, but feeding back the CSI to the base station is problematic due to the inevitable data-rate limitation on the feedback link. This motivates the research into limited feedback systems where only partial CSI is fed back to the base station [2], [3].

Another challenging problem of multi-user schemes in general is scheduling. The scheduling algorithm should have a low complexity, but the transmissions to the users must still fulfill strict Quality-of-Service (QoS) constraints. An important QoS constraint is the minimum signal to interference-plus-noise ratio (SINR). The minimum SINR constraint requires that every scheduled user in the cell has an SINR larger than a predefined threshold. Further, especially for modern real-time multimedia communication systems, it is important that the delay between two transmissions to the same user remains constant, i.e., that the users are scheduled in a round-robin fashion. Another advantage of round-robin scheduling is the reduced overhead since the base station does not need to sacrifice transmission time to inform the users in every block about their allocated slot positions. However, due to the stochastic nature of the wireless channel, it is not possible to provide hard QoS guarantees, i.e., if the channel is in a deep fade it is not possible to fulfill the SINR constraint.

The large sum rate on the vector Gaussian broadcast channel is achievable with dirty paper coding (DPC) [1], [4], [5]. DPC has a high computational complexity, but order-optimal performance is also possible with zero-forcing beamforming in the high SNR regime [6]. In [7], a scheduling algorithm for zero-forcing beamforming was proposed that takes the individual queue lengths at the base station into account. However, zero-forcing beamforming requires perfect CSI at the base station. One of the first schemes to exploit the multiuser diversity assuming a data-rate limited feedback link is orthogonal beamforming (OB). OB was presented in [8] using opportunistic scheduling to maximize the instantaneous sum rate. The price of using opportunistic scheduling is the lack of short-term fairness, i.e., fairness is only achieved in the long run. A low-complexity scheduling algorithm for OB was proposed in [9]. The effect of partial CSI at the base station on zero-forcing beamforming was investigated in [10]. In [11] the authors investigate the benefits of feeding back the SINR together with the quantized direction of the zero-forcing beamforming vectors.

Another important aspect of beamforming is the distribution of the available power over the different beamforming vectors. The solution to the power allocation for maximizing the minimum SINR of the scheduled user was presented in [12], and a solution to fulfill individual SINR constraints on the users was presented in [13]. The concept of using the limited CSI feedback for beamforming is also used in the PU2RC algorithm [14], [15], an algorithm used in recent systems, e.g., IEEE 802.16m.

An algorithm that takes the time-varying nature of the channel into account, but still provides strong bounds on the maximum delay, is the Channel Aware Round-Robin (CARR) scheduling algorithm [16]. It schedules every user once inside each block, but it does not implement true round-robin scheduling, since the positions of the users inside the block are dynamically allocated. The CARR algorithm chooses the positions depending on the channel state of the different users in the different slots. Thus, the maximum delay between two transmissions using the CARR algorithm is two block lengths.

Manuscript received November 26, 2009; revised October 22, 2010 and May 11, 2011; accepted May 26, 2011. The associate editor coordinating the review of this paper and approving it for publication was D. Gesbert.

This research was supported in part by NWO-STW under the VIDI program (DTC.6577).

The material in this paper has been presented in part at the IEEE Workshop Signal Process. Advances Wireless Commun. (SPAWC), Perugia, Italy, June 2009.

C. Simon can be contacted at clasimon@mac.com.

G. Leus is with the Delft University of Technology, The Netherlands.

Digital Object Identifier 10.1109/TWC.2011.062011.091736

The main disadvantage of CARR scheduling compared to round-robin scheduling is the additional overhead. For CARR scheduling the base station has to inform the users for every block in what slot they are scheduled. A similar scheduling algorithm for space division multiple access (SDMA) is the Best Fit algorithm [17]. The Best Fit algorithm also tries to assign all the users in every block, but it uses SDMA to dynamically assign multiple users to the same slot depending on the resulting SINR. It further considers an SINR constraint. A low-complexity variant is the Partial Best Fit (PBF) algorithm [18]. It just adds new users according to the Best Fit strategy and removes the expired users, i.e., the users that have no more packets to transmit. An overview of other algorithms that consider scheduling under the exploitation of the spatial diversity can be found in [19].

Our proposed algorithm tries to schedule all the users in the cell in a round-robin fashion as long as possible. The application of orthogonal beamforming allows to reduce the interference between users scheduled at the same time instant. Further, it also reduces the feedback requirements from the users to the base station since full CSI feedback is not necessary. We propose a corresponding feedback metric, and we consider the necessary quantization due to the data-rate limited feedback link. The feedback is used by the base station to dynamically divide the available transmit power among the users. This allows the weakest users, i.e., the users with the worst channel conditions, to fulfill the SINR constraint longer than with an equal power distribution. A user is rescheduled if he is no longer able to fulfill the SINR constraint despite receiving additional power, i.e., the user is scheduled at a different slot and with a different beamformer in the next block. The performance of the algorithm is depicted through simulations for a time-varying channel.

Notation: We use capital boldface letters to denote matrices, e.g., $\mathbf{A}$, and small boldface letters to denote vectors, e.g., $\mathbf{a}$. $E(\cdot)$ denotes expectation, and $P(\cdot)$ probability. We will denote the probability density function (pdf) of the random variable X as $f_X(x)$, and the cumulative distribution function (cdf) as $F_X(x)$. We write the logical conjunction between two values x and y as x and y, and the logical disjunction as x or y.

II. System Model

We assume a narrowband single-cell scenario where a base station with M antennas transmits data to N single-antenna users. At a given time user i receives the symbol

$$y_i = \sum_{j \in \mathcal{S}} \mathbf{h}_i \mathbf{w}_{g(j)} \sqrt{P_j} s_j + n_i \tag{1}$$

where $\mathcal{S}$ contains the indices of the users scheduled at that time instant, $\mathbf{h}_i \in \mathbb{C}^{1 \times M}$ is the channel of user i, and $\mathbf{w}_{g(j)} \in \mathbb{C}^{M \times 1}$ is the beamforming vector assigned to user j. The mapping $g(j)$ maps a beamforming vector from the same beamformer codebook $\mathcal{W}$ to every user.

The power assigned to user j is denoted P_j, and the data symbol s_j, that is transmitted to user j, is selected from a constellation with average unit power. The noise n_i is complex Gaussian distributed with zero mean and variance N_0, i.e., $n_i \sim \mathcal{CN}(0, N_0)$. The total allocated transmit power is limited to $P_T = \sum_{i \in \mathcal{S}} P_i$, and the signal-to-noise ratio (SNR) of the system is $\text{SNR} = \frac{P_T}{N_0}$.

The users have the possibility to feed back information to the base station at the start of every block. The feedback link itself is instantaneous, error-free, and data-rate limited to B bits. All the users have to fulfill a strict SINR constraint denoted SINR_{min}. We assume that the individual users acquire perfect channel state information (CSI) at the start of each block through training. The time-correlated channel is modeled according to Jakes' model [20].

Every block consists of K slots. We further assume that the channel is block-fading, i.e., the channel is constant throughout the K slots of a block. The block index k starts at $k = 0$, and the slot index l restarts at the beginning of each new block at $l = 0$. Thus, the relation between the current time instant t and the current block/slot index is $t = kK + l$.

We are using a set of orthogonal beamforming vectors from a codebook $\mathcal{W}$ to simultaneously transmit to maximally M users [8]. The codebook $\mathcal{W}$ contains M orthogonal beamforming vectors $\mathbf{w}_m$. The M beamformers in the codebook all have unit norm, i.e., $\|\mathbf{w}_m\|_2 = 1, m \in \mathcal{M} = \{1, \ldots, M\}$. The codebook $\mathcal{W}$ is known to the users and to the base station. Note that a possible extension would be to consider multiple orthogonal beamforming codebooks.

The main objective is to schedule the users in a round-robin fashion. If a user i has been scheduled at time instant $t = (k-1)K + l$ using the beamformer $\mathbf{w}_m$, then we want to schedule it also at time instant $t = kK + l$ using the beamformer $\mathbf{w}_m$. Further, all the scheduled users have to fulfill a strict SINR constraint, i.e., they need to have an SINR higher than SINR_{min}.

Serving the users in a round-robin fashion should result in a packet delay variation of zero. However, due to the time-varying nature of the wireless channel, there is a non-zero probability that the channel is in a deep fade, i.e., reliable communication is not possible. Thus, it is not possible to guarantee the QoS constraints, i.e., to have hard QoS guarantees. The problem is now to exploit the available feedback link to schedule the users as long as possible in a round-robin fashion while still fulfilling the SINR constraint.

III. Algorithm Overview

We assume that the different users are able to acquire perfect CSI at the beginning of each block, i.e., $t = kK, \forall k$. Due to the block-fading nature of the channel, an individual user thus has perfect channel knowledge for every slot in the block. Using this channel knowledge the user then calculates how much power the base station has to assign to it in order to reach the SINR constraint.

Next, this minimum power is quantized and fed back to the base station. Once the base station receives all the feedback from the users, it checks for every time slot if the sum of the fed back quantized minimum powers exceeds the maximally allocatable transmit power P_T at the base station. If the sum is lower, then all the users in that slot can be scheduled using the available transmit power. However, if the sum is higher then it is not possible to schedule all the users. The users with the highest power demands are dropped until the sum of the required power for the remaining users is lower than the available transmit power.

In the next step, the dropped users from the previous block and the users who just entered the cell are scheduled. Once all the users have been assigned to a slot, their required power is assigned to them and the remaining power is equally distributed to all the users in that slot.

A. Feeding Back the Required Power

The SINR for user i is calculated as

$$\mathrm{SINR}_i = \frac{|\mathbf{h}_i \mathbf{w}_{g(i)}|^2 P_i}{\sum_{j \in \mathcal{S}\setminus\{i\}} |\mathbf{h}_i \mathbf{w}_{g(j)}|^2 P_j + N_0}. \quad (2)$$

We see that the SINR of user i depends on the individual transmit powers of the users in the set $\mathcal{S}$. In order to determine the minimum amount of power that is required by user i to reach $\mathrm{SINR}_{\min}$, it is necessary to know the amount of power that is assigned to the other users scheduled in the same slot. However, the individual power levels assigned to the other users in the set are not known to the individual users. A solution is to feed back the full CSI to the base station and to balance the SINR between the different users using the algorithm in [12]. The drawback is that it requires full channel knowledge or at least knowledge of the composite channel energies $|\mathbf{h}_i \mathbf{w}_{g(j)}|^2, j \in \mathcal{S}$ at the base station and thus incorporates a lot of feedback.

In this paper, we try to find an estimate of the power assigned to user i that fulfills the SINR constraint, and that does not depend on the power levels assigned to the other users in the set $\mathcal{S}$. This required power will be denoted $\hat{P}_i$. We start by defining an estimate of the true SINR, denoted $\hat{\mathrm{SINR}}_i$, that does not depend on how the total transmit power is distributed over the users in $\mathcal{S}$, but that is guaranteed to be smaller than the true SINR

$$\hat{\mathrm{SINR}}_i \leq \mathrm{SINR}_i. \quad (3)$$

Due to (3), it is certain that if the estimated SINR fulfills the SINR constraint, so does the true SINR, i.e., if $\mathrm{SINR}_{\min} \leq \hat{\mathrm{SINR}}_i$ then $\mathrm{SINR}_{\min} \leq \mathrm{SINR}_i$. We propose to use

$$\hat{\mathrm{SINR}}_i = \frac{|\mathbf{h}_i \mathbf{w}_{g(i)}|^2 P_i}{\max_{j \in \mathcal{M}\setminus\{g(i)\}} |\mathbf{h}_i \mathbf{w}_j|^2 (P_T - P_i) + N_0} \quad (4)$$

which is lower than or equal to the real SINR since inserting (4) and (2) into (3) results in

$$\sum_{j \in \mathcal{S}\setminus\{i\}} |\mathbf{h}_i \mathbf{w}_{g(j)}|^2 P_j \leq \max_{j \in \mathcal{M}\setminus\{g(i)\}} |\mathbf{h}_i \mathbf{w}_j|^2 (P_T - P_i) \quad (5)$$

which is always true. The minimum power assigned to a user i that fulfills the SINR constraint and that only depends on the total power can thus be calculated from (4) as

$$\hat{P}_i = \frac{\max_{j \in \mathcal{M}\setminus\{g(i)\}} |\mathbf{h}_i \mathbf{w}_j|^2 P_T + N_0}{\frac{1}{\hat{\mathrm{SINR}}_{\min}} |\mathbf{h}_i \mathbf{w}_{g(i)}|^2 + \max_{j \in \mathcal{M}\setminus\{g(i)\}} |\mathbf{h}_i \mathbf{w}_j|^2}. \quad (6)$$

Next, the power $\hat{P}_i$ is quantized and fed back to the base station.

The probability distribution of the required power depends on whether the user is scheduled for the first time or not. If a user is scheduled for the first time, then it chooses the beamforming vector that maximizes its SINR, i.e., the user chooses the beamforming vector

$$g(i) := \arg\max_{j \in \mathcal{M}} |\mathbf{h}_i \mathbf{w}_j|^2. \quad (7)$$

For the successive blocks, the beamforming vector $g(i)$ might no longer be the beamforming vector that results in the highest SINR for user i. However, even with this suboptimal beamforming vector, the user i might fulfill the SINR constraint.

As mentioned, if a user is scheduled for the first time, then its beamforming vector is determined using (7). For that case we call the required power the initial required power. The cdf and the pdf of the initial required power are derived in Appendix B. For every successive scheduling instant, however, the true channel changes according to the assumed channel model, but the selected beamforming vector remains the same. In order to simplify the derivation of the cdf and the pdf of the required power, we assume, just for the derivation, that the channel is i.i.d. between the scheduling instances. This corresponds to a scenario where the channel has a high Doppler spread or where the user has been scheduled for a long time in the same slot. We call the resulting required power, the regular required power. The cdf and the pdf of the regular required power are derived in Appendix A.

B. Quantizing the Feedback

The data rate limitation on the feedback link makes a quantization of the minimum power necessary before it can be fed back to the base station. The quantization Q maps the minimum power $\hat{P}_i$ to an element of a predefined codebook $\mathcal{C} = \{c_1, \ldots, c_b\}$, i.e., $Q : \mathbb{R}^+ \to \mathcal{C}$. We assume that the codebook size is limited to $b = 2^B$ entries in order to fulfill the data-rate limitation of the feedback link. The required power $\hat{P}_i$ of user i is quantized using

$$Q(\hat{P}_i) = \arg\min_{c_q \in \mathcal{C}} \; c_q - \hat{P}_i \quad \text{s.t.} \quad \hat{P}_i \leq c_q \quad (8)$$

and the index q of the element $c_q = Q(\hat{P}_i)$ of the codebook $\mathcal{C}$ is fed back to the base station. The quantized minimum power of user i is denoted $\hat{P}_{Q,i} = Q(\hat{P}_i)$.

In order to prevent that a user gets too little power assigned due to the quantization error, the condition $\hat{P}_i \leq Q(\hat{P}_i)$ has to be fulfilled, and therefore we must define $c_b = +\infty$. We further define $c_{b-1} = P_T$. However, this might not be optimal in the sense of maximizing the number of scheduled users. This is best visualized by imagining the case of having only 1 bit available to quantize the required power. Then, using the previous reasoning, we would use the codebook $\mathcal{C}_1 = \{P_T, +\infty\}$. The disadvantage of $\mathcal{C}_1$ is that the users can only be scheduled in a TDMA fashion, since every user requests either all the available transmit power P_T, or he declares that he cannot be scheduled. If we would use the codebook $\mathcal{C}_2 = \{\frac{P_T}{M}, +\infty\}$, then only users would be scheduled that reach the SINR constraint assuming an equal power allocation and $M-1$ interfering users. The disadvantage of codebook $\mathcal{C}_2$ is that if user i has a required power between $\frac{P_T}{M} < \hat{P}_i \leq P_T$ he cannot be scheduled, whereas that user could have been scheduled with codebook $\mathcal{C}_1$. On the other hand, the use of $\mathcal{C}_2$ allows the scheduling of up to M users simultaneously. We see that the number of users in the cell and also the number of available slots inside a block must be taken into account when a codebook is designed that maximizes the average number of scheduled users. However, in order to keep the problem tractable, we assume that the number of users in the cell remains small enough so that the TDMA mode is beneficial, i.e., we always take $c_{b-1} = P_T$.

C. Codebook Design

In order to simplify the notation we will substitute $\hat{P}_i$ with x throughout this section. Designing the codebook requires the definition of a distortion metric $d(x, Q(x))$ which serves as a measure for the quality of the quantization. The most popular metric used in the quantization literature is the mean squared error [21]. However, for quantizing the required power, the absolute error is a better metric, since it corresponds to minimizing the overall power loss due to the quantization. Using this metric, the average distortion D of a codebook $\mathcal{C}$ is calculated as

$$D(\mathcal{C}) = \int_0^{+\infty} |x - Q(x)| f_P(x)dx \tag{9}$$

where $f_P(x)$ is the pdf of the required power. Inserting the selection function (8), we can rewrite (9) as

$$D(\mathcal{C}) = \sum_{q=1}^{b} \int_{c_{q-1}}^{c_q} (c_q - x) f_P(x)dx \tag{10}$$

with $c_0 = 0$ since the required power of a user is always positive. We see that, compared to classic quantizer design, the codebook elements and the regions are directly linked. However, fixing $c_b = +\infty$ makes (10) ill-defined since the occurence of an element in the region $(P_T, +\infty)$ leads to an infinite average distortion. However, since c_{b-1} and c_b are already fixed, we can also restrict ourselves to minimizing the simplified distortion function

$$D_s(\mathcal{C}) = \sum_{i=1}^{b-1} \int_{c_{i-1}}^{c_i} (c_i - x) f_P(x)dx \tag{11}$$

with $c_0 = 0$ and $c_{b-1} = P_T$. We start by rewriting (11) as

$$D_s(\mathcal{C}) = \sum_{i=1}^{b-1} \int_{c_{i-1}}^{c_i} c_i f_P(x)dx - k \tag{12}$$

with $k = \int_0^{P_T} x f_P(x)dx$. It is possible to show through simulations that $D_s(\mathcal{C})$ is not convex and not quasiconvex. Thus, we cannot solve the problem directly using standard tools. We start by looking for the codebooks that are critical points of the simplified distortion function D_s. The gradient of the distortion function is zero for the critical points of the distortion function

$$\nabla D_s(\mathcal{C}) = \mathbf{0}_{b-2} \tag{13}$$

where $\mathbf{0}_{b-2}$ is a $(b-2)$-dimensional column vector with all entries being 0. Since c_{b-1} is determined beforehand we just have $b - 2$ variables in our problem. The critical points are found by solving

$$\frac{\partial}{\partial c_i} \sum_{j=1}^{b-1} \int_{c_{j-1}}^{c_j} c_j f_P(x)dx = 0 \tag{14}$$

for $i = 1, \ldots, b-2$, which can be simplified to

$$\frac{\partial}{\partial c_i} \int_{c_{i-1}}^{c_i} c_i f_P(x)dx + \frac{\partial}{\partial c_i} \int_{c_i}^{c_{i+1}} c_{i+1} f_P(x)dx = 0 \tag{15}$$

for $i = 1, \ldots, b-2$. We use the Leibniz Integral Rule to solve (15) and obtain

$$\int_{c_{i-1}}^{c_i} f_P(x)dx - (c_{i+1} - c_i) f_P(c_i) = 0 \tag{16}$$

for $i = 1, \ldots, b-2$. A codebook $\mathcal{C}$ that fulfills (16) is a critical point. It is not possible to design the codebook based on the Lloyd's Method 1 [22] since there are no iterative optimality conditions to solve, i.e., nearest neighbor condition and centroid condition. However, it is possible to design the codebook using variational techniques, e.g., Lloyd's Method 2 [22]. The basic idea is to solve (16) by fixing c_1 and then to calculate the remaining elements c_i for $i = 2, \ldots, b-2$ as

$$c_{i+1} = \frac{1}{f_P(c_i)} \int_{c_{i-1}}^{c_i} f_P(x)dx + c_i \tag{17}$$

until finally we have c_{b-1}. We assumed initially that c_{b-1} should be P_T. However, the resulting c_{b-1} provided by (17) might not result in $c_{b-1} = P_T$. Thus, if $h(c_1) = c_{b-1} - P_T$ is negative (positive), then it means that c_1 was chosen too small (large). Then we choose a larger (smaller) c_1 until we finally find $h(c_1) = 0$ and thus $c_{b-1} = P_T$. Using the results from Appendices A and B, the problem can be easily solved numerically.

Note that we implicitly assume that the codeword c_{b-1} is a strictly monotonically increasing function of c_1. Numerical simulations show that this holds for the investigated cases. The solution is then also assumed to be the unique solution of (11), although we can not rigorously prove this.

IV. Scheduling

After the base station receives the instantaneous feedback from all users, it schedules the users for the current block k using the following three steps.

1) Validation of the Scheduled Users: The base station starts by assigning for every slot $l = 0, \ldots, K-1$ in the current block k the same users that were scheduled in the same slot in the previous block, i.e., $\mathcal{S}[kK + l] := \mathcal{S}[(k-1)K + l]$, using the same beamforming vectors as in the previous block. Then, the base station calculates the sum of the required powers for every slot using the fed back required energies. If the sum is larger than the available transmit power P_T then it is not possible to schedule the users so that they all fulfill the SINR constraint while still using the same slot and the same beamformer as in the last block. The straightforward solution is to remove the user from the set that has the highest power demand, i.e., the user that feeds back the highest required power. The dropped user is added to the set $\mathcal{U}_{\text{resched}}[kK]$ and will be treated in the next block. This is repeated until the sum of the minimum powers of the remaining users is smaller than the available transmit power P_T. This first step of the scheduling algorithm is described in Algorithm 1.

Algorithm 1 Validation of the Scheduled Users

$\mathcal{U}_{\text{resched}}[kK] := \emptyset$
for $l = 0$ to $K-1$ **do**
  $t := kK + l$
  $\mathcal{S}[t] := \mathcal{S}[t-K]$
  while $\sum_{i\in\mathcal{S}[t]} \hat{P}_{Q,i}[kK] > P_T$ **do**
    $i := \arg\max_{i\in\mathcal{S}[t]} \hat{P}_{Q,i}[kK]$
    $\mathcal{S}[t] := \mathcal{S}[t] \setminus \{i\}$
    $\mathcal{U}_{\text{resched}}[kK] := \mathcal{U}_{\text{resched}}[kK] \cup \{i\}$
  end while
end for

2) Scheduling the New Users and Rescheduling the Dropped Users: We collect the dropped users from the previous block $\mathcal{U}_{\text{resched}}[(k-1)K]$ and the new users entering the cell $\mathcal{U}_{\text{new}}[kK]$ in the set $\mathcal{U}[kK] := \mathcal{U}_{\text{resched}}[(k-1)K] \cup \mathcal{U}_{\text{new}}[kK]$. All these users feed back their required power as well as the index of the corresponding beamforming vector. Thus, user $i \in \mathcal{U}[kK]$ feeds back $g(i)$ from (7) and $\hat{P}_{Q,i}[kK]$. If we assume that the data-rate limitation on the feedback link is strict, then we have to use a codebook with $2^{B-\lceil \log_2 M\rceil}$ entries to quantize $\hat{P}_i[kK]$. The base station then tries to successively schedule all the users in $\mathcal{U}[kK]$ according to their fed back minimum power. The algorithm starts by scheduling the user with the largest power requirement first. Once this user has been found, the base station looks for a slot that is not yet using the preferred beamformer of the considered user. For every one of these free slots the base station calculates the sum of the minimum power levels of the users in the slot, assuming the considered user is added, and finally chooses the slot that results in the lowest sum. If the base station does not find a valid slot then the user is skipped for the current block and added to $\mathcal{U}_{\text{resched}}[kK]$. However, if the base station finds a slot then the user is scheduled for transmission. This second step of the scheduling algorithm is described in Algorithm 2.

Algorithm 2 Scheduling the New Users and Rescheduling the Dropped Users

$\mathcal{U} := \mathcal{U}_{\text{resched}}[(k-1)K] \cup \mathcal{U}_{\text{new}}[kK]$
while $|\mathcal{U}| > 0$ **do**
  $i := \arg\max_{i\in\mathcal{U}} \hat{P}_{Q,i}[kK]$
  Temp_Index $:= -1$
  Temp_Power $:= P_T$
  for $l = 0$ to $K-1$ **do**
    if $\sum_{j\in\mathcal{S}[kK+l]} \hat{P}_{Q,j}[kK] + \hat{P}_{Q,i}[kK] \leq$ Temp_Power **and** $g(i) \neq g(j), \forall j \in \mathcal{S}[kK+l]$ **then**
      Temp_Index $:= l$
      Temp_Power $:= \sum_{j\in\mathcal{S}[kK+l]} P_{Q,j}[kK] + \hat{P}_{Q,i}[kK]$
    end if
  end for
  if Temp_Index $\neq -1$ **then**
    $\mathcal{S}[kK+\text{Temp\_Index}] := \mathcal{S}[kK+\text{Temp\_Index}] \cup \{i\}$
  else
    $\mathcal{U}_{\text{resched}}[kK] := \mathcal{U}_{\text{resched}}[kK] \cup \{i\}$
  end if
  $\mathcal{U} := \mathcal{U} \setminus \{i\}$
end while

3) Power Assignment: Once all the users are scheduled, every scheduled user is assigned its required minimum power. The remaining transmit power is uniformly distributed over the remaining users in the same slot. Thus, the transmit power for user i, that is allocated to $t = kK + l$, is calculated as

$$P_i[t] := \hat{P}_{Q,i}[kK] + \frac{P_T - \sum_{j\in\mathcal{S}[t]} \hat{P}_{Q,j}[kK]}{|\mathcal{S}[t]|}. \quad (18)$$

Note that this approach tries to balance the SINRs of the different users in the same slot. However, due to the quantization, and due to the unknown interference between the users in the set $\mathcal{S}[t]$, it is not possible to truly balance the SINRs as it is possible with full CSI at the base station [12]. It is also possible to save power at the base station by solely assigning the required power to the users and not redistributing the spare power. Then, (18) simply becomes

$$P_i[t] := \hat{P}_{Q,i}[kK]. \quad (19)$$

V. Simulations

We start by comparing the performance of the codebook design algorithm from Section III-C to different common codebook design strategies. We design two codebooks using the algorithm from Section III-C. The first codebook, denoted the regular codebook, is designed using the statistics from Appendix A, and the second codebook, denoted the initial codebook, is designed using the statistics from Appendix B. Their performance is compared with a uniform codebook and with an equiprobable codebook. All the codebooks contain the values P_T and $+\infty$ as their two largest elements. Thus, we can only freely choose the remaining $b-2$ elements from the codebooks. The ith element of the uniform codebook is calculated as $i\frac{P_T}{b-1}$ with $i = 1, \ldots, b-2$. The equiprobable codebook is designed such that all the elements of the codebook are selected with the same probability. First calculating $P_{\text{avg}} = \frac{P(0\leq \hat{P}_i \leq P_T)}{b-1}$, the different elements are successively calculated by solving $F_X(c_i) - F_X(c_{i-1}) = P_{\text{avg}}$ where $c_0 = 0$ for $i = 1, \ldots, b-2$. For the simulations presented in Fig. 1 and Fig. 2, we consider the average quantization error per user as a function of the codebook size. The simulations consider an average SNR of 10 dB, $M = 4$, and a minimum required SINR of 7 dB. The Rayleigh channel model was used for both simulations. We also created some simple test cases to investigate the optimality of the proposed codebook design algorithm. In all these test cases, the critical point turned out to be the optimal point.

We see in Fig. 1 that the initial codebook has a lower quantization error $E(\hat{P}_{Q,i} - \hat{P}_i)$ than the other codebooks for quantizing the power immediately after scheduling. This is expected since the initial codebook is designed using the proper statistics for the first scheduling instance from Appendix A. Further, we see in Fig. 1 that the regular codebook, which is designed using the mismatched statistics, performs as good as the equiprobable codebook that is designed using the correct statistics, at least for small codebook sizes. However, with every successive block the assumption that the user uses its optimal beamformer, i.e., the beamforming vector that

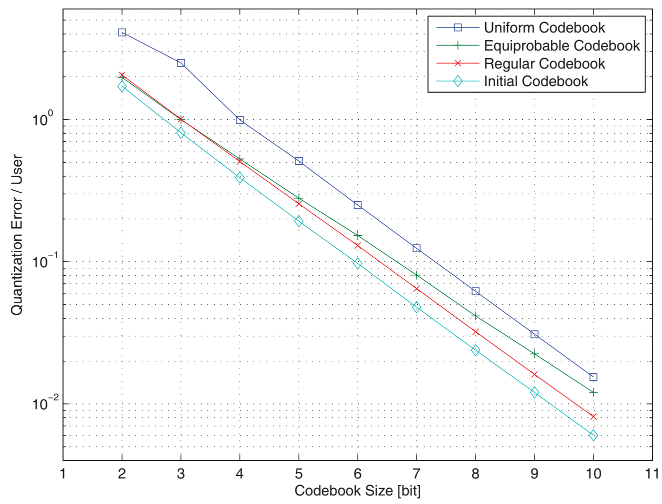

Fig. 1. Comparison of different codebook design approaches for scheduling a user at the first slot. ($M = 4$, $|\mathcal{S}| = 4$, SNR $= 10$ dB, $\text{SINR}_{\text{min}} = 7$ dB).

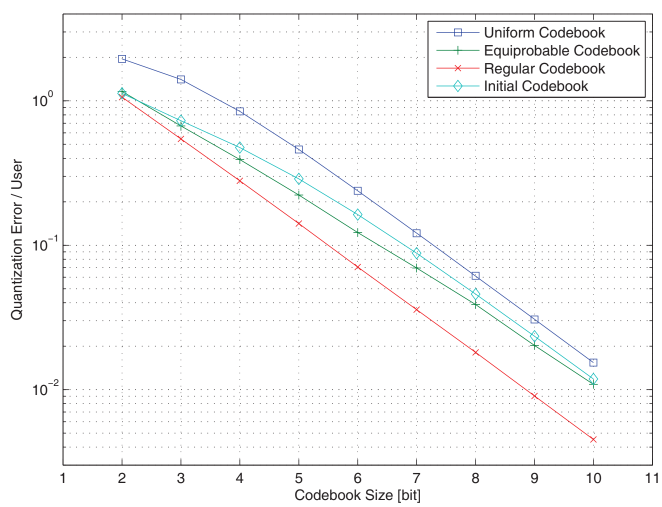

Fig. 2. Comparison of different codebook design approaches for scheduling a user at a later slot. ($M = 4$, $S = 4$, SNR $= 10$ dB, $\text{SINR}_{\text{min}} = 7$ dB).

maximizes the SINR, becomes weaker.

The extreme case is depicted in Fig. 2. Here we assume that the beamforming vector is selected randomly amongst all the available beamforming vectors. This corresponds to the case where the user is able to fulfill its SINR constraint successively over a prolonged time using the same beamformer. We see that the regular codebook performs well for both cases, and thus, for the sake of simplicity, we use the regular codebook for the following simulations. Note that a codebook switching strategy, where we use the initial codebook for the first slot, and then switch to the regular codebook should provide a minor performance gain, but requires more storage capacity from the users and from the base station.

The simulation depicted in Fig. 3 shows how long the different users are successively scheduled on the average as a function of the product of the Doppler frequency f_{D} and the block length T_{f}. We assume a homogeneous cell where all the users experience SNR $= 15$ dB. The SINR constraint is fixed to $\text{SINR}_{\text{min}} = 5$ dB. The time-correlation between the blocks is modeled according to Jakes' model. At time instant t the

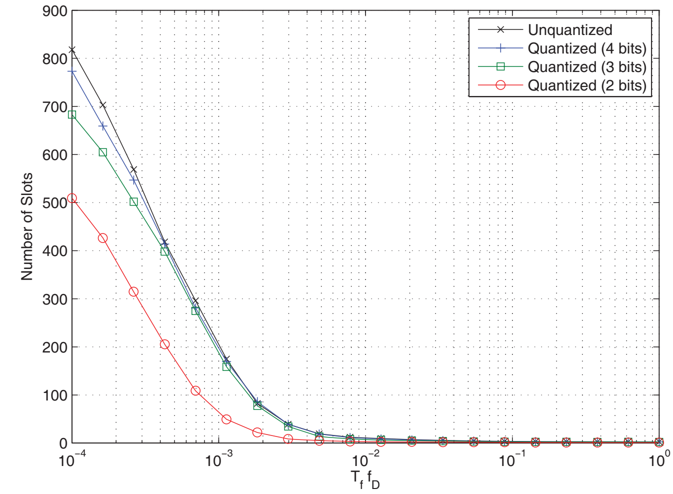

Fig. 3. Average number of slots that a user is successively scheduled for a varying product of block length T_{f} and Doppler frequency f_{D}. ($M = 2$, $K = 100$, 150 users, 1000 blocks, SNR $= 15$ dB, $\text{SINR}_{\text{min}} = 5$ dB).

pth element from the channel $\mathbf{h}_i[t]$ is modelled as

$$[\mathbf{h}_i[t]]_p = \frac{1}{\sqrt{Q}} \sum_{q=1}^{Q} a_{p,q} \exp(j\, 2\pi T_{\text{f}} f_{\text{D}} \lceil t/K \rceil \cos \alpha_{p,q}) \quad (20)$$

where Q is the number of scatterers, $a_{p,q}$ is i.i.d. complex Gaussian distributed with zero mean and variance 1, and $\alpha_{p,q}$ is uniformly distributed over $[0, 2\pi]$. We assume $Q = 30$ scatterers. The influence of quantizing the fed back minimum power is depicted for multiple codebooks and for no quantization. We see that for slowly changing channels, i.e., channels with a low product of Doppler frequency f_{D} and block length duration T_{f}, the average number of consecutive blocks increases. We simulate 1000 blocks for every channel realization, and thus the maximum number of blocks a user can be successively scheduled is limited to 1000. However, if $f_{\text{D}} T_{\text{f}}$ increases, then the channel becomes more volatile. This increases $P(\sum_{i \in \mathcal{S}[t-K]} \hat{P}_{Q,i}[t] > P_T \mid \sum_{i \in \mathcal{S}[t-K]} \hat{P}_{Q,i}[t - K] \leq P_T)$, i.e., the probability that the users from the set $\mathcal{S}[t-K]$ in the slot $t-K$ cannot fulfill the SINR constraint in the slot t and thus have to be rescheduled.

Fig. 4 depicts the average number of users that are scheduled per block as a function of $T_{\text{f}} f_{\text{D}}$. We see that the average number of scheduled users decreases as the channel becomes more volatile. This is because the probability that a user has to be rescheduled increases with the volatility of the channel, and every rescheduled user is not scheduled for at least one block. We also see the effects of quantizing the required power. A smaller codebook size leads to a larger power loss since the user requests more power than he actually needs, due to the quantization error. For small codebooks, the system nearly operates exclusively in TDMA mode, i.e., every slot has just a single user assigned to.

The next simulation, Fig. 5, compares the feedback of the required power to having full CSI at the base station. We assume that in both cases Algorithm 2 is used to schedule the new and the dropped users. However, the base station with full CSI uses the results from [12] to calculate the optimal power assigned to every beamforming vector to balance the resulting

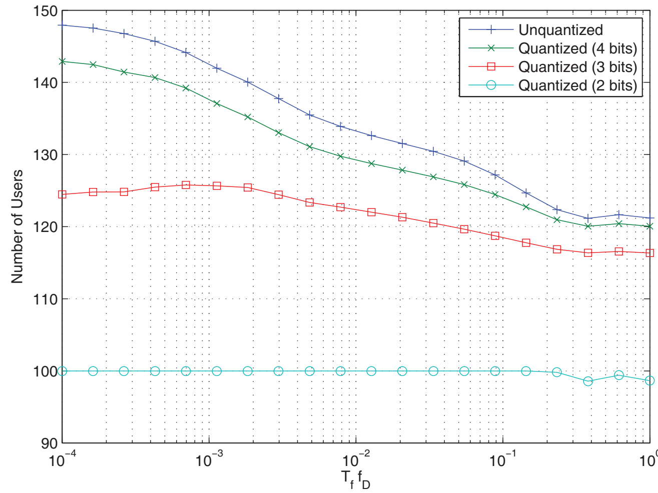

Fig. 4. Average number of users that are scheduled per block for a varying product of block length T_{f} and Doppler frequency f_{D}. ($M = 2$, $K = 100$, 150 users, 1000 blocks, $Q = 30$, SNR $= 15$ dB, $\mathrm{SINR}_{\mathrm{min}} = 5$ dB).

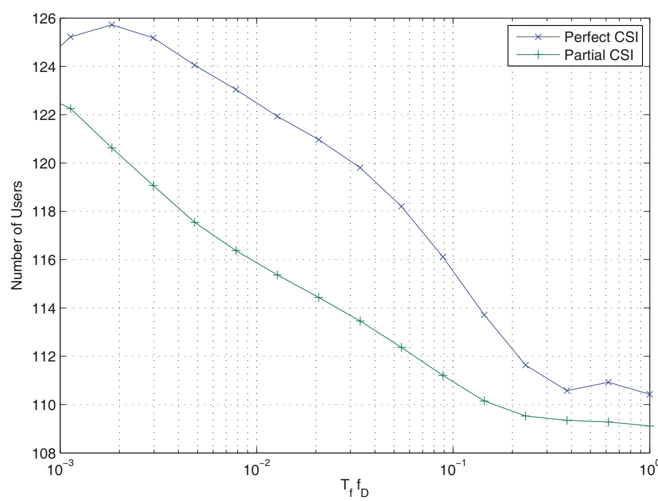

Fig. 5. Average number of users that are scheduled in the same slot as in the previous slot for a varying product of block length T_{f} and Doppler frequency f_{D}. ($M = 3$, $K = 100$, 150 users, 1000 blocks, $Q = 30$, SNR $= 15$ dB, $\mathrm{SINR}_{\mathrm{min}} = 5$ dB).

SINRs. We see that having full CSI allows to schedule more users successively in the same slot.

Fig. 6 shows the effect of erroneous CSI on the scheduling. We create the noisy channel estimate $\mathbf{h}_{i,\mathrm{noisy}}$ by adding noise to the true channel, i.e., $\mathbf{h}_{i,\mathrm{noisy}} = \mathbf{h}_{i,\mathrm{true}} + \mathbf{e}_i$ with $\mathbf{e}_i \in \mathbb{C}^{1\times M}$. The different components of the noise vector are complex Gaussian distributed with zero mean and variance N_E. The users calculate their required power based on the noisy channel, and feed it back to the base station. Then, the base station uses the feedback to schedule the users. Next, we check, using the true channel, how many of the scheduled users really fulfill the SINR constraint.

VI. Conclusions

We presented a scheme to implement round-robin scheduling using orthogonal beamforming and data-rate limited feedback. The scheme uses scalar feedback from the users to divide the transmit power amongst the users so that they all

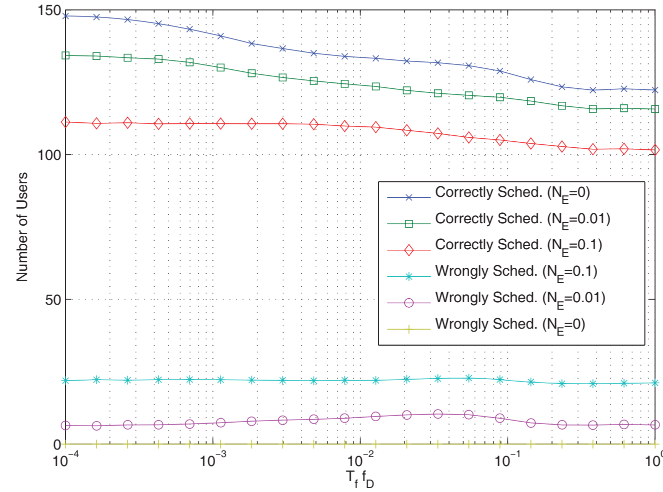

Fig. 6. The effect of erroneous CSI on the number of correctly and wrongly scheduled users in the slot for a varying product of block length T_{f} and Doppler frequency f_{D}. ($M = 2$, $K = 100$, 150 users, 1000 blocks, $Q = 30$, SNR $= 15$ dB, $\mathrm{SINR}_{\mathrm{min}} = 5$ dB).

fulfill a given SINR constraint. The simulations show that the presented algorithm is attractive to implement round-robin scheduling for time-varying channels. We further propose an algorithm to design the codebooks used to quantize the feedback, and the codebooks outperform other popular codebooks.

Appendix A
CDF and PDF of the Regular Required Power

We assume that the channel $\mathbf{h}_i$ of user i is a random variable. Thus, the resulting required power $\hat{P}_i$ is also a random variable, and we want to derive the corresponding cdf

$$F_P(z) = P(\hat{P}_i \leq z)$$
$$= P\left(\frac{\max_{j\in\mathcal{M}\setminus g(i)} |\mathbf{h}_i\mathbf{w}_j|^2 P_T + N_0}{\frac{1}{\widehat{\mathrm{SINR}}_{\mathrm{min}}}|\mathbf{h}_i\mathbf{w}_{g(i)}|^2 + \max_{j\in\mathcal{M}\setminus g(i)} |\mathbf{h}_i\mathbf{w}_j|^2} \leq z\right) \tag{21}$$

and pdf $f_P(z) = \frac{d}{dz}F_P(z)$. We assume in this section that the function $g : \{1,\ldots,K\} \to \mathcal{M}$ randomly assigns a beamforming vector to a user. The set of orthogonal beamforming vectors are unit-norm $\|\mathbf{w}_m\|_2 = 1$, $\forall m$ and known. The different elements of the channel $\mathbf{h}_i$ are assumed i.i.d. and circular Gaussian distributed according to $\mathcal{CN}(0,1)$. In order to simplify the notation we will write (21) as

$$F_P(z) = P\left(\frac{X_A P_T + N_0}{X_B + X_A} \leq z\right) \tag{22}$$

where $X_A = \max_{j\in\mathcal{M}\setminus g(i)} |\mathbf{h}_i\mathbf{w}_j|^2$, and $X_B = \frac{1}{\widehat{\mathrm{SINR}}_{\mathrm{min}}}|\mathbf{h}_i\mathbf{w}_{g(i)}|^2$. The real part and the imaginary part of $\mathbf{h}_i\mathbf{w}_j$ are independently Gaussian distributed $\mathcal{N}(0,\frac{1}{2})$. We define the continuous random variable $X_1 = |\mathbf{h}_i\mathbf{w}_j|^2$, and model this variable as $X_1 = \frac{1}{2}(X_{11}^2 + X_{12}^2)$, with $X_{11} \sim \mathcal{N}(0,1)$ and $X_{12} \sim \mathcal{N}(0,1)$. The term $X_{11}^2 + X_{12}^2$ is $\chi^2(2)$ distributed. The pdf $f_{X_1}(x)$ of X_1 then is

$$f_{X_1}(x) = \begin{cases} e^{-x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

and the corresponding cdf $F_{X_1}(x)$ is

$$F_{X_1}(x) = \begin{cases} 1 - e^{-x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}. \quad (24)$$

The different realizations of $|\mathbf{h}_i \mathbf{w}_j|^2$ are identical and independently distributed for all $j \in \mathcal{M}$. The cdf of $X_A = \max_{j \in \mathcal{M} \setminus g(i)} |\mathbf{h}_i \mathbf{w}_j|^2$ then is

$$F_{X_A}(x) = (F_{X_1}(x))^{M-1} = \begin{cases} (1 - e^{-x})^{M-1} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}. \quad (25)$$

The resulting pdf of $X_A = \max_{j \in \mathcal{M} \setminus i} |\mathbf{h}_i \mathbf{w}_j|^2$ is

$$f_{X_A}(x) = \begin{cases} e^{-x}(1 - e^{-x})^{M-2}(M-1) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}. \quad (26)$$

The pdf of $X_B = \frac{1}{\hat{\text{SINR}}_{\text{min}}} |\mathbf{h}_i \mathbf{w}_{g(i)}|^2$ corresponds to $X_B = \frac{1}{\hat{\text{SINR}}_{\text{min}}} X_1$, and is

$$f_{X_B}(x) = \begin{cases} \hat{\text{SINR}}_{\text{min}} \, e^{-x \, \hat{\text{SINR}}_{\text{min}}} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}. \quad (27)$$

Since the random variables X_A and X_B are always positive, we can rewrite (22) as

$$F_P(z) = P\left((P_T - z)X_A - zX_B \leq -N_0\right). \quad (28)$$

The random variable $(P_T - z)X_A - zX_B$ will be abbreviated as X_C in order to keep the notation compact. We use (26) to find the pdf of the random variable $(P_T - z)X_A$ which is

$$f_{(P_T - z)X_A}(x) = \frac{M-1}{|P_T - z|} e^{-\frac{x}{P_T - z}} (1 - e^{-\frac{x}{P_T - z}})^{M-2} \quad (29)$$

if $((x \geq 0 \text{ and } P_T \geq z) \text{ or } (x \leq 0 \text{ and } P_T \leq z))$ and $f_{(P_T - z)X_A}(x) = 0$ otherwise. The pdf of $-zX_B$ is found by using (27), and is

$$f_{-zX_B}(x) = \frac{\hat{\text{SINR}}_{\text{min}}}{|z|} \, e^{\frac{x}{z} \hat{\text{SINR}}_{\text{min}}} \quad (30)$$

if $((x \geq 0 \text{ and } z \leq 0) \text{ or } (x \leq 0 \text{ and } z \geq 0))$ and $f_{-zX_B}(x) = 0$ otherwise. Since $(P_T - z)X_A$ and $-zX_B$ are statistically independent, the pdf f_{X_C} can be calculated as

$$f_{X_C}(x) = \int_{-\infty}^{+\infty} f_{(P_T - z)X_A}(x - y) f_{-zX_B}(y) dy. \quad (31)$$

The product $f_{(P_T - z)X_A}(x - y) f_{-zX_B}(y)$ is not zero when $[(x \geq y \text{ and } P_T \geq z) \text{ or } (x \leq y \text{ and } P_T \leq z)]$ and $[(y \geq 0 \text{ and } z \leq 0) \text{ or } (y \leq 0 \text{ and } z \geq 0)]$. Using Boolean algebra this can be rewritten as $(P_T \leq z \text{ and } x \leq y \text{ and } y \leq 0 \text{ and } z \geq 0)$ or $(P_T \leq z \text{ and } x \leq y \text{ and } y \geq 0 \text{ and } z \leq 0)$ or $(P_T \geq z \text{ and } x \geq y \text{ and } y \leq 0 \text{ and } z \geq 0)$ or $(P_T \geq z \text{ and } x \geq y \text{ and } y \geq 0 \text{ and } z \leq 0)$. We can thus distinguish between the following four cases:

- Case 1: $P_T \leq z$ and $x \leq y$ and $y \leq 0$ and $z \geq 0$
- Case 2: $P_T \leq z$ and $x \leq y$ and $y \geq 0$ and $z \leq 0$
- Case 3: $P_T \geq z$ and $x \geq y$ and $y \leq 0$ and $z \geq 0$
- Case 4: $P_T \geq z$ and $x \geq y$ and $y \geq 0$ and $z \leq 0$

We see in (22) that $F_P(z) = 0, \forall z < 0$. Thus, we can ignore Case 2 and Case 4. Next, we rewrite (31) as

$$f_{X_C}(x) = \begin{cases} f_{X_C,1}(x) & 0 \leq z \leq P_T \\ f_{X_C,2}(x) & P_T \leq z \\ 0 & \text{otherwise} \end{cases} \quad (32)$$

with

$$f_{X_C,1}(x) = \frac{(M-1)\hat{\text{SINR}}_{\text{min}}}{|P_T - z| z} e^{-\frac{x}{P_T - z}} \int_{-\infty}^{\min(0,x)} e^{y\left(\frac{1}{P_T - z} + \frac{\hat{\text{SINR}}_{\text{min}}}{z}\right)} (1 - e^{-\frac{x-y}{P_T - z}})^{M-2} dy. \quad (33)$$

Since we assume that $M \leq 2$ we can rewrite the previous equation as

$$f_{X_C,1}(x) = \frac{(M-1)\hat{\text{SINR}}_{\text{min}}}{|P_T - z| z} e^{-\frac{x}{P_T - z}} \sum_{k=0}^{M-2} \binom{M-2}{k} (-1)^k e^{-\frac{kx}{P_T - z}} \int_{-\infty}^{\min(0,x)} e^{y\left(\frac{k+1}{P_T - z} + \frac{\hat{\text{SINR}}_{\text{min}}}{z}\right)} dy. \quad (34)$$

Similarly, the second case in (32) becomes

$$f_{X_C,2}(x) = \frac{(M-1)\hat{\text{SINR}}_{\text{min}}}{|P_T - z| z} e^{-\frac{x}{P_T - z}} \sum_{k=0}^{M-2} \binom{M-2}{k} (-1)^k e^{-\frac{kx}{P_T - z}} \int_{x}^{0} e^{y\left(\frac{k+1}{P_T - z} + \frac{\hat{\text{SINR}}_{\text{min}}}{z}\right)} dy. \quad (35)$$

Using f_{X_C} we can now calculate the cdf

$$F_P(z) = \begin{cases} F_{P,1}(z) = \int_{-\infty}^{-N_0} f_{X_C,1}(x) dx & 0 \leq z \leq P_T \\ F_{P,2}(z) = \int_{-\infty}^{-N_0} f_{X_C,2}(x) dx & P_T \leq z \\ 0 & \text{otherwise} \end{cases}. \quad (36)$$

On the region $0 \leq z \leq P_T$ the cdf is given by

$$F_{P,1}(z) = \sum_{k=0}^{M-2} \binom{M-2}{k} (-1)^k \frac{(M-1)z}{zk + z + \hat{\text{SINR}}_{\text{min}}(P_T - z)} e^{-N_0 \frac{\hat{\text{SINR}}_{\text{min}}}{z}}. \quad (37)$$

The cdf $F_{P,1}(z)$ can be written in a more compact form. We start by rewriting (37) as

$$F_{P,1}(z) = (M-1)! \sum_{k=0}^{M-2} \frac{(-1)^k}{(M-2-k)!k!} \frac{(-1)^k}{(k + 1 + \hat{\text{SINR}}_{\text{min}}(\frac{P_T}{z} - 1))} e^{-N_0 \frac{\hat{\text{SINR}}_{\text{min}}}{z}} \quad (38)$$

Using Lemma 1 we can rewrite it as

$$F_{P,1}(z) = \Gamma(M) \frac{\Gamma(1 + \hat{\text{SINR}}_{\text{min}}(\frac{P_T}{z} - 1))}{\Gamma(M + \hat{\text{SINR}}_{\text{min}}(\frac{P_T}{z} - 1))} e^{-N_0 \frac{\hat{\text{SINR}}_{\text{min}}}{z}}. \quad (39)$$

On the region $P_T < z$ the cdf is given by

$$F_{P,2}(z) = -\sum_{k=0}^{M-2} \binom{M-2}{k} (-1)^k \frac{(M-1)\hat{\text{SINR}}_{\text{min}}}{zk + z + \hat{\text{SINR}}_{\text{min}}(P_T - z)} \left(\frac{P_T - z}{-1-k} e^{-N_0 \frac{-1-k}{P_T - z}} - \frac{z}{\hat{\text{SINR}}_{\text{min}}} e^{-N_0 \frac{\hat{\text{SINR}}_{\text{min}}}{z}} \right). \quad (40)$$

Using (36) the pdf of the required power can now be calculated using simple derivation.

Lemma 1: Let $a \in \mathbb{R}$ with $x > 0$, and $b \in \mathbb{N}$. Then, the identity

$$\sum_{k=0}^{b} \frac{(-1)^k}{k!(b-k)!(k+1+a)} = \prod_{j=1}^{b+1} \frac{1}{a+j} \quad (41)$$

is true.

Proof: We prove the lemma by induction. It is easy to see that the equation is valid for the induction basis, i.e., $b = 0$. Next, we have to show that (41) is true for b if it is true for $b - 1$. We start by rewriting (41) as

$$\sum_{k=0}^{b} \frac{(-1)^k}{k!(b-k)!(k+1+a)} = \frac{1}{a+b+1} \prod_{j=1}^{b} \frac{1}{a+j}. \quad (42)$$

Now, we insert the induction hypothesis and obtain

$$\sum_{k=0}^{b} \frac{(-1)^k}{k!(b-k)!(k+1+a)} = \frac{1}{a+b+1} \sum_{k=0}^{b-1} \frac{(-1)^k}{k!(b-1-k)!(k+1+a)} \quad (43)$$

$$\Rightarrow \sum_{k=0}^{b} \frac{(-1)^k}{k!(b-k)!} = 0. \quad (44)$$

Using the identity

$$\sum_{k=0}^{b} (-1)^k \binom{b}{k} = 0 \quad (45)$$

from [23], we see that (44) is true. Thus, the induction step is proved as well. ■

APPENDIX B
CDF AND PDF OF THE INITIAL REQUIRED POWER

As in Appendix A we want to calculate the cdf of the initial required power

$$F_P(z) = P(\hat{P}_{\text{min}} \leq z) = P\left(\frac{\max_{j \in \mathcal{M} \setminus g(i)} |\mathbf{h}_i \mathbf{w}_j|^2 P_T + N_0}{\frac{1}{\hat{\text{SINR}}_{\text{min}}} |\mathbf{h}_i \mathbf{w}_{g(i)}|^2 + \max_{j \in \mathcal{M} \setminus g(i)} |\mathbf{h}_i \mathbf{w}_j|^2} \leq z \right) \quad (46)$$

and its corresponding pdf $f_P(z) = \frac{d}{dz} F_P(z)$. However, in contrast to Appendix A we assume that $g(i) = \arg\max_{j \in \mathcal{M}} |\mathbf{h}_i \mathbf{w}_j|^2$. Thus, we cannot assume that $\max_{j \in \mathcal{M} \setminus g(i)} |\mathbf{h}_i \mathbf{w}_j|^2$ and $|\mathbf{h}_i \mathbf{w}_{g(i)}|^2$ are statistically independent anymore. We define again the continuous random variable $X_1 = |\mathbf{h}_i \mathbf{w}_j|^2$, with the pdf (23), and the cdf (24). Using ordering statistics notation, we write $X_{(M-1)} = \max_{j \in \mathcal{M} \setminus g(i)} |\mathbf{h}_i \mathbf{w}_j|^2$ and $X_{(M)} = |\mathbf{h}_i \mathbf{w}_{g(i)}|^2$ with $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(M)}$, and thus

$$F_P(z) = P\left(\frac{X_{(M-1)} P_T + N_0}{\frac{1}{\hat{\text{SINR}}_{\text{min}}} X_{(M)} + X_{(M-1)}} \leq z \right). \quad (47)$$

The problem can be reformulated as

$$F_P(z) = P((X_{(M-1)}, X_{(M)}) \in D_z) \quad (48)$$
$$= \iint_{D_z} f_j(x, y)\, dx\, dy \quad (49)$$

where the region D_z is

$$D_z = \{(x, y) \in \mathbb{R}^2 \mid \frac{x P_T + N_0}{\frac{1}{\hat{\text{SINR}}_{\text{min}}} y + x} \leq z\} \quad (50)$$

and the joint density f of $X_{(M-1)}$ and $X_{(M)}$ is calculating using [24, Eq. (2.1.6)] as

$$f(x, y) = \begin{cases} \frac{M!}{(M-2)!} F_{X_1}[x]^{M-2} f_{X_1}(y) f_{X_1}(x) & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}. \quad (51)$$

However, since (23), (24), and (51) are piecewise functions, the evaluation of (49) is cumbersome. Thus, we define the region $D_1 \subset D_z$

$$D_1 := \{(x, y) \in D_z \mid 0 \leq x \text{ and } 0 \leq y \text{ and } x \leq y\} \quad (52)$$

where f is non-zero. Next, we define the function $f_1 : D_1 \to \mathbb{R}$ as

$$f_1(x, y) = \frac{M!}{(M-2)!} (1 - e^{-y})^{M-2} e^{-y} e^{-x}. \quad (53)$$

Then, using (52) and (53), we can rewrite (49) as

$$F_P(z) = \iint_{D_z} f(x, y)\, dx\, dy = \iint_{D_1} f_1(x, y)\, dx\, dy \quad (54)$$

The domain D_1 can now be calculated as

$$D_1 = \begin{cases} D_2 & \text{if } P_T - z > 0 \\ D_3 & \text{if } P_T - z < 0 \end{cases} \quad (55)$$

$$D_2 = \{(x, y) \in \mathbb{R} \mid 0 < x < ay - b \text{ and } 0 < x \leq y\} \quad (56)$$

$$D_3 = \{(x, y) \in \mathbb{R} \mid ay - b < x \text{ and } 0 < x \leq y \text{ and } 0 \leq y\} \quad (57)$$

with $a = \frac{z}{(P_T - z)\hat{\text{SINR}}_{\text{min}}}$ and $b = \frac{N_0}{P_T - z}$. We start by calculating D_2, i.e., we assume that $P_T - z > 0$. We see that D_2 is bounded by the line $L_1 = \{(x, y) \in \mathbb{R}^2 \mid x = ay - b\}$ and $L_2 = \{(x, y) \in \mathbb{R}^2 \mid x = y\}$. We differentiate between the following three cases:

- $a > 1 \Leftrightarrow z > \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1 + \hat{\text{SINR}}_{\text{min}}}$
- $a = 1 \Leftrightarrow z = \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1 + \hat{\text{SINR}}_{\text{min}}}$
- $a < 1 \Leftrightarrow z < \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1 + \hat{\text{SINR}}_{\text{min}}}$

Line L_1 and L_2 intersect at a point with the exception of the second case, i.e., $a = 1$, where L_1 and L_2 are parallel. The intersection point (x_c, y_c) is calculated as

$$x_c = y_c = -\frac{N_0 \hat{\text{SINR}}_{\text{min}}}{\hat{\text{SINR}}_{\text{min}}(P_T - z) - z}. \quad (58)$$

Another point of interest is the intersection of L_1 with the y-axis. This point, denoted $(0, y_0)$, is calculated as

$$y_0 = \frac{N_0 \hat{\text{SINR}}_{\text{min}}}{z}. \quad (59)$$

We start by investigating $a > 1$. We assume that L_1 and L_2 intersect on the line segment $\{(x, x) \in \mathbb{R}^2 \mid x < 0\}$, i.e., we assume that

$$-\frac{N_0 \hat{\text{SINR}}_{\text{min}}}{\hat{\text{SINR}}_{\text{min}}(P_T - z) - z} < 0 \quad (60)$$

$$\Rightarrow \quad \frac{P_T \hat{\text{SINR}}_{\text{min}}}{\hat{\text{SINR}}_{\text{min}} + 1} > z \quad (61)$$

which contradicts that $a > 1$, and thus L_1 and L_2 must intersect on $\{(x, x) \in \mathbb{R}^2 \mid x \geq 0\}$. We can now define the domain D_2 for $\frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}} < z < P_T$ as

$$D_2 = \{(x, y) \in \mathbb{R} \mid (0 < x < ay - b \text{ and } y_0 < y < y_c) \text{ or } (0 < x < y \text{ and } y_c < y))\}. \quad (62)$$

Next, we investigate $a = 1$, i.e., L_1 and L_2 are parallel. Since $b > 0$, the border of D_2 is defined by L_1, and L_1 intersects with the y-axis at y_0. We can then define the domain D_2 for $z = \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}}$ as

$$D_2 = \{(x, y) \in \mathbb{R} \mid 0 < x < ay - b \text{ and } y_0 < y)\}. \quad (63)$$

Finally, we assume that $a < 1$. We approach this case as the $a > 1$ case: We can prove that the intersection of L_1 and L_2 has negative coordinates, and next, we determine the intersection of L_1 and the y-axis. We can then define the domain D_2 for $0 < z < \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}}$ as

$$D_2 = \{(x, y) \in \mathbb{R} \mid 0 < x < ay - b \text{ and } y_0 < y\}. \quad (64)$$

We see that we can omit to treat the $a = 1$ case separately by including it with $a > 1$ or $a < 1$. We choose $a > 1$, and thus $a \geq 1$ covers the region $\frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}} \leq z < P_T$. Finally, we can express the region D_2 as

$$D_2 = \{(x, y) \in \mathbb{R} \mid (0 < x < ay - b \text{ and } y_0 < y)\} \quad (65)$$

if $0 < z < \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}}$ and

$$D_2 = \{(x, y) \in \mathbb{R} \mid (0 < x < ay - b \text{ and } y_0 < y < y_c) \text{ or } (0 < x < y \text{ and } y_c < y))\} \quad (66)$$

if $\frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}} \leq z < P_T$. We can now proceed to determine the region D_3. Using the fact that $z > P_T$ and using (61) we see that the intersection point of L_1 and L_2 has positive coordinates. Since $a < 0$, the region D_3 is

$$D_3 = \{(x, y) \in \mathbb{R} \mid (ay - b < x < y \text{ and } y_c < y < y_0) \text{ or } (0 < x < y \text{ and } y_0 < y)\} \quad (67)$$

and thus all the domaines are determined. The resulting cdf is

$$F_P(z) = \begin{cases} 0 & z < 0 \\ F_{P1}(z) & 0 < z < \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}} \\ F_{P2}(z) & \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}} < z < P_T \\ F_{P3}(z) & P_T < z \end{cases}. \quad (68)$$

The different pieces of the cdf can now be determined using straightforward integration. On the domain $(0, \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}})$ the cdf $F_P(z)$ is

$$F_{P1}(z) = \int_{y_0}^{+\infty} \int_0^{ay-b} f_1(x, y)\, dx\, dy \quad (69)$$

$$F_{P1}(z) = \sum_{k=0}^{M-2} (-1)^k \frac{M!}{k!(M-2-k)!} \frac{z}{(P_T - z)\hat{\text{SINR}}_{\text{min}} + z(k+1)} e^{-\frac{N_0 \hat{\text{SINR}}_{\text{min}}}{z}}. \quad (70)$$

On the domain $[\frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}}, P_T)$, $F_P(z)$ is calculated as

$$F_{P2}(z) = \int_{y_0}^{y_c} \int_0^{ay-b} f_1(x, y)\, dx\, dy + \int_{y_c}^{+\infty} \int_0^{y} f_1(x, y)\, dx\, dy \quad (71)$$

and results in (72). Finally, on the domain $[P_T, +\infty)$, the cdf F_P is determined by evaluating

$$F_{P3}(z) = \int_{y_c}^{y_0} \int_{ay-b}^{y} f_1(x, y)\, dx\, dy + \int_{y_0}^{+\infty} \int_0^{y} f_1(x, y)\, dx\, dy. \quad (73)$$

We see that $F_{P3} = F_{P2}$, and thus the cdf is now

$$F_P(z) = \begin{cases} 0 & z < 0 \\ F_{P1}(z) & 0 < z < \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}} \\ F_{P2}(z) & \frac{P_T \hat{\text{SINR}}_{\text{min}}}{1+\hat{\text{SINR}}_{\text{min}}} < z \end{cases}. \quad (74)$$

The corresponding pdf can now be calculated by straightforward derivation.

References

[1] P. Viswanath and D. N. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[2] D. J. Love, R. W. Heath Jr., V. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[3] D. Gesbert, M. Kountouris, R. W. Heath Jr., C.-B. Chae, and T. Sälzer, "Shifting the MIMO paradigm," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 36–46, Sep. 2007.

[4] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, Sep. 2006.

[5] M. H. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[6] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

$$F_{P2}(z) = \sum_{k=0}^{M-2} \frac{M!}{k!(M-2-k)!(-k-1)}(-1)^k \left(\left(\frac{(P_T - z)\widehat{\mathrm{SINR}}_{\min}}{(-k-1)z - (P_T - z)\widehat{\mathrm{SINR}}_{\min}} + \frac{1}{k+2} \right) e^{\frac{N_0 \widehat{\mathrm{SINR}}_{\min}(k+2)}{\widehat{\mathrm{SINR}}_{\min}(P_T - z) - z}} - \left(\frac{(P_T - z)\widehat{\mathrm{SINR}}_{\min}}{(-k-1)z - (P_T - z)\widehat{\mathrm{SINR}}_{\min}} + 1 \right) e^{-\frac{N_0 \widehat{\mathrm{SINR}}_{\min}}{z}} \right) \quad (72)$$

[7] C. Swannack, E. Uysal-Biyikoglu, and G. W. Wornell, "Low complexity multiuser scheduling for maximizing throughput in the MIMO broadcast channel," in *Proc. Allerton Conf. on Comm. Control and Comp.*, Sep. 2004.

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

[9] R. de Francisco, M. Kountouris, D. T. Slock, and D. Gesbert, "Orthogonal linear beamforming in MIMO broadcast channels," in *Proc. IEEE Wireless Communications and Networking Conference*, Mar. 2007, pp. 1210–1215.

[10] C. Swannack, E. Uysal-Biyikoglu, and G. W. Wornell, "MIMO broadcast scheduling with limited channel state information," in *Proc. Allerton Conference on Communication, Control, and Computing*, Sep. 2005.

[11] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna broadcast channels with limited feedback and user selection," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1478–1491, Sep. 2007.

[12] W. Yang and G. Xu, "Optimal downlink power assignment for smart antenna systems," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 6, Apr. 1998, pp. 3337–3340.

[13] M. Schubert and H. Boche, "Solution of the multiuser downlink beamforming problem with individual SINR constraints," *IEEE Trans. Veh. Commun.*, vol. 53, no. 1, pp. 18–28, Jan. 2004.

[14] S.-J. Kim, K.-H. Kim, C. S. Park, and K. B. Lee, "Mobile communication apparatus and method including base station and mobile station having multiple antenna," Patent No. US 7,324,480 B2, Jan. 2008.

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

[16] H.-Y. Wei and R. Izmailov, "Channel-aware soft bandwidth guarantee scheduling for wireless packet access," in *Proc. IEEE Wireless Communications and Networking Conference*, vol. 2, Mar. 2004, pp. 1276–1281.

[17] F. Shad, T. D. Tod, V. Kezys, and J. Litva, "Dynamic slot allocation (DSA) in indoor SDMA/TDMA using a smart antenna basestation," *IEEE/ACM Trans. Netw.*, vol. 9, no. 1, pp. 69–81, Feb. 2001.

[18] H. Yin and H. Liu, "Performance of space-division multiple-access (SDMA) with scheduling," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 611–618, Oct. 2002.

[19] W. Ajib and D. Haccoun, "An overview of scheduling algorithms in MIMO-based fourth-generation wireless systems," *IEEE Netw.*, vol. 19, no. 5, pp. 43–48, Sep. 2005.

[20] W. C. Jakes, Jr., *Microwave Mobile Communications*. John Wiley & Sons, 1974.

[21] Z. Wang and A. C. Bovik, "Mean squared error: love it or leave it?" *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[22] S. P. Lloyd, "Least squares quantization in PCM," Bell Laboratories, Whippky, NJ, United States, Tech. Rep., July 1957.

[23] M. Abramowitz and I. A. Stegun, editors, *Handbook of Mathematical Functions*. Dover, 1972.

[24] A. E. Sarhan and B. G. Greenberg, editors, *Contributions to Order Statistics*. John Wiley & Sons, Inc., 1962.

Claude Simon received a degree in electrical engineering (Dipl.-Ing.) from the Technical University of Kaiserlautern, Germany, in 2005. He worked from 2005 to 2010 as a research assistant at the Delft University of Technology. His current research interests are vector quantization and wireless communication with limited feedback.

Geert Leus was born in Leuven, Belgium, in 1973. He received the electrical engineering degree and the PhD degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in June 1996 and May 2000, respectively. He has been a Research Assistant and a Postdoctoral Fellow of the Fund for Scientific Research - Flanders, Belgium, from October 1996 till September 2003. During that period, Geert Leus was affiliated with the Electrical Engineering Department of the Katholieke Universiteit Leuven, Belgium. Currently, Geert Leus is an Associate Professor at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology, The Netherlands. During the summer of 1998, he visited Stanford University, and from March 2001 till May 2002 he was a Visiting Researcher and Lecturer at the University of Minnesota. His research interests are in the area of signal processing for communications. Geert Leus received a 2002 IEEE Signal Processing Society Young Author Best Paper Award and a 2005 IEEE Signal Processing Society Best Paper Award. He was the Chair of the IEEE Signal Processing for Communications and Networking Technical Committee, and an Associate Editor for the IEEE Transactions on Signal Processing, the IEEE Transactions on Wireless Communications, and the IEEE Signal Processing Letters. Currently, he serves on the Editorial Board of the EURASIP Journal on Applied Signal Processing.