Makki and Eriksson *EURASIP Journal on Wireless Communications and Networking* 2013, **2013**:249
http://jwcn.eurasipjournals.com/content/2013/1/249

EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Fairness, power allocation, and CSI quantization in block fading multiuser systems

Behrooz Makki* and Thomas Eriksson

Abstract

This paper studies the fairness, power allocation, and channel state information (CSI) quantization in multiuser systems utilizing multiple feedback bits per user. For a given set of schedulers, we obtain the system throughput with different power allocation strategies and any combination of different fading distributions. Moreover, the system performance under user different outage probability constraints is investigated. Assuming homogenous users, as a special case, the throughput is determined for fixed and *random request networks*. Considering nonidentical fading channels between the transmitter and receivers, the throughput is found under the *K-significant average* feedback bit allocation technique, and two suboptimal fairness schemes are investigated which satisfy different quality-of-service requirements. The results show that using optimal power allocation, the first quantization region (QR) of each user is the only QR for which no power may be allocated. The system outage probability vanishes as the number of users goes to infinity. Users hard outage probability constraints can be satisfied at the cost of one more QR in the channel quantizer. Finally, the proposed fairness schemes are more flexible than the standard proportional fair (PF) scheduling in dealing with the throughput-fairness tradeoff. However, their superiority over PF scheduling depends on the fairness constraint.

1 Introduction

1.1 Motivation

It has been demonstrated both practically [1-8] and theoretically [9-19] that employment of adaptive modulation and scheduling leads to substantial performance improvement in multiuser systems, normally called *multiuser diversity*. Traditionally, the fading is considered as an unreliability source which should be mitigated. In the multiuser diversity context, however, the channel fading has a positive impact and is helpful for improving the system performance [1-19]. This is because in a system with a number of users experiencing independent fading conditions, it is more likely that at each time instant, one of the users experiences *good* channel quality. Hence, the data transmission efficiency is improved by always communicating the *best* users.

In order to prioritize among the users and select the proper modulation for the best user, the scheduler must, in theory, know the channels perfectly which, due to feedback signaling overhead, is not practically feasible. Hence, a quantized representation of the channel state information (CSI), expressed via a limited number of feedback bits, is normally provided at the transmitters. This is a simple and practical approach motivating the current limited-feedback schemes such as [1-8] and this paper.

In a limited-feedback multiuser system, different aspects should be taken into account, among which fairness [9,15-17] is one of the most important ones. As discussed in [20], the users in different locations of a cellular coverage area are subject to distance-dependent path loss that may have more than 30-dB dynamic range. That is, in realistic propagation conditions, the channels from a transmitter to the users are not identically distributed. Hence, the rate and power optimization in terms of throughput, as the sole constraint, leads to giving all resources to the users close to the transmitter, while leaving users at the cell edges to starve. To avoid this problem, fairness is an additive constraint widely studied in the literature. However, considering the fairness as a constraint results in sacrificing part of the throughput, to provide equality between the users. Therefore, it is important to study the throughput-fairness tradeoff in multiuser systems.

*Correspondence: behrooz.makki@chalmers.se
Department of Signals and Systems, Chalmers University of Technology, Hörsalsvägen 11, 41269 Gothenburg, Sweden

1.2 Literature review

Assuming different levels of CSI, many scientific reports can be found that have tackled the multiuser diversity problem in different theoretical and practical aspects. For instance, [9-11] studied the performance of multiuser networks under perfect CSI assumption. Considering limited feedback, [2,3] have presented reviews of the related papers published until 2008. In the following, a description of the recently published papers, most relevant to our work, is presented[a].

Most of the proposed limited-feedback multiuser models are based on one-bit feedback and threshold-based schemes, where the CSI is fed back when it exceeds a threshold. In these methods, the system performance has been mostly investigated with fixed transmission power and for homogeneous networks, i.e., networks with independent and identically distributed (i.i.d) channels, which require no fairness discussion. For example, a threshold-based scheme was proposed by [21], where the users inform the transmitter about their channels' quality only if it passes a threshold. Almost the same results were obtained in [22], where multiuser diversity was exploited in slowly fading random access channels. Also, [8] and [12] studied the effect of one-bit feedback and a per-user threshold scheme on the performance of multiuser networks, respectively. With one-bit feedback, [23] found the throughput of multiple-input multiple-output (MIMO) broadcast systems. Moreover, in [24], the users report their CSI via dedicated feedback channels if their scheduling metric is greater than a threshold. Finally, considering erroneous delayed quantized CSI feedback, [25] found the average throughput of i.i.d multiuser orthogonal frequency division with multiple access (OFDMA) systems utilizing M-level quadrature amplitude modulation (M-QAM) or M-level phase shift keying (M-PSK) modulations.

Multiple feedback bits per user has been addressed in a number of recent works, where the quantized CSI is obtained via random vector quantization or predefined tables. For instance, utilizing M-QAM or binary PSK techniques, [14,19,26-28] studied the spectral efficiency of i.i.d multiuser systems in the presence of multiple-bit quantized CSI. Here, the quantizers are determined based on given tables determined by the desired bit error rates. Then, [29] investigated different feedback quantization schemes in multiuser networks where the quantizers are designed such that the probability of not selecting the user with the best channel quality is minimized. Also, [30] analyzed the throughput scaling of a per-user unitary and rate control scheme for an asymptotically large number of users. Here, the *gain shape* is quantized by randomly generated feedback codebooks, while the signal-to-noise-and-interference ratio (SINR) of each user is assumed to be perfectly known at the transmitter. Moreover, using random vector quantization, [31] found the ergodic achievable rates of a multimode multiuser MIMO channel in the presence of delayed quantized CSI feedback. Khoshnevis and Yu [32] studied different feedback bit allocation approaches in multiuser systems using a shared feedback channel. The feedback bits are distributed such that the average transmission power is minimized, subject to individual users' outage probability. In all mentioned papers, the results are obtained with fixed transmission power in i.i.d channels. Finally, considering the quantized CSI in orthogonal multiple access channels, [33] studied the weighted average power minimization subject to individual users' minimum average rate constraint. Here, the optimal transmission parameters are obtained for given quantization boundaries.

In contrast to the mentioned papers, there are a number of schemes developed for fairness in limited-feedback multiuser systems. For instance, [34] proposed a threshold-based scheduling approach to tackle the capacity and fairness tradeoff in multiuser systems. Here, the users are divided into two groups. In the first group, the users whose normalized signal-to-noise ratio (SNR) exceeds a threshold send feedback. In the second group, a number of the best remaining users whose absolute SNR passes the threshold send one-bit feedback. The results of [34] were extended in [17], where cases in which none of the users' channel quality exceeds the threshold were specially considered. Furthermore, [35,36] studied the capacity and fairness in multiuser diversity systems under a specific feedback scheme. Here, the users share a common feedback channel period which is divided into some subperiods. In each subperiod, if a user channel quality passes threshold(s), it sends some feedback and, if it is the only user sending feedback, the feedback process ends. If two or more users send feedback simultaneously, the feedback signal is irrecoverable and lost. Finally, considering uniform power allocation, [37] used quantized CSI to maximize the outage capacity under a long-term fairness constraint. The fairness is obtained by (1) clustering the users and (2) determining the quantization regions (QRs) in each cluster, such that the probability that equal number of users fall into a region is the same.

1.3 Contributions

In this paper, we study the throughput, power allocation, fairness, and CSI quantization schemes in a scheduler-equipped multiuser system using multiple feedback bits per user. The goal is to maximize the throughput subject to different transmission power or fairness constraints. Therefore, the main challenge is to determine the optimal quantization boundaries, transmission rates, and powers, such that the system data transmission efficiency is optimized under different quality-of-service requirements.

Particularly, the paper demonstrates (1) the effect of adaptive power allocation and user activation probability on the throughput of the multiuser systems utilizing quantized CSI feedback, (2) what optimality conditions the quantization parameters should satisfy and how they are affected by increasing the power and number of users, (3) how different long-term or semi-instantaneous fairness constraints can be satisfied by combination of different suboptimal scheduling policies, (4) how a user outage probability affects the system throughput, and (5) what is the throughput of a system utilizing the K-significant average scheme for feedback bit allocation.

In comparison to the previous literature, the new points considered in the paper are as follows: (1) The number of feedback bits is not limited to one, and the users can have different number of feedback bits. (2) As opposed to all reviewed papers, except [33] where the goal is to minimize the average power in orthogonal multiple access channels, the throughput is obtained under both optimal and uniform power allocation strategies. (3) In contrast to some other works, e.g., [23,33], the data outage probability, due to imperfect CSI at the transmitter, is considered in the throughput calculation. Also, the system performance under users' hard outage probability constraints is studied. (4) Both long-term and semi-instantaneous fairness constraints are addressed with practical scheduling schemes. The fairness schemes are different from the ones considered in [17,34-37]. Finally, (5) the results are not restricted to specific schedulers, and the effect of user activation probability and K-significant average feedback bit allocation on the throughput is discussed. Thus, the paper is a generalization of what have been done in, e.g., [8,12,21,22], and new discussions on the quantizer optimality conditions, random request networks, optimal power allocation, user outage probability, and fairness are presented which, to the best of authors knowledge, have not been investigated before.

The main conclusions of the paper are as follows: (1) optimal channel quantization affects the system throughput substantially, specifically when the number of users increases or the transmission power decreases. (2) With uniform power allocation, the throughput of a user with extremely hard outage probability constraint and M channel QRs is the same as the throughput with no outage probability constraint and $M-1$ QRs. (3) The system outage probability vanishes as the number of users goes to infinity. (4) Using optimal power allocation, the first QR of each user is the only region for which no power may be allocated. (5) For a large range of scheduling policies, the maximum (un)fair throughput is achieved when the channel is assumed to be its worst value within each QR, except the first one. (6) The optimal QRs expand as the number of users increases. Also, the QRs get closer to origin when the power increases. Finally, (7) the proposed fairness schemes are more flexible than the standard proportional fair (PF) scheduling model in dealing with the throughput-fairness tradeoff. However, their superiority over PF scheduling depends on the fairness constraint.

1.4 An overview of the paper

The system model and the notations are presented in Section 2. Power-limited throughput optimization problem in the case when there is no fairness constraint is first studied in Section 3. Although this is a special case of the general problem, it provides a basis that can be used when the fairness constraints are added. Demonstrating the general problem, the throughput is obtained under uniform and optimal power allocation strategies, which is then followed by discussions about the optimality conditions of the quantization parameters. Finally, we study two special cases, where fixed or random number of users experience the same fading pdfs, and when the *k-significant average technique* [4-7] is implemented for feedback bit allocation.

The throughput-fairness tradeoff is addressed in Section 4. First, the system throughput is investigated in the case, where each individual user is constrained to have a given minimum throughput over long time. We call this condition as *long-term fairness* constraint. Here, due to the nonconvexity of the optimization problem, a numerical algorithm is developed which can be implemented in different optimization problems.

Although long-term fairness constraint guarantees the user average performance, there may be cases where they remain off in many successive time slots. However, depending on the application, this may not be desirable for the users. Hence, we consider a *semi-instantaneous fairness* constraint which is defined as the condition that the users are served by the transmitter not always but within limited time slots. Here, using the round robin approach, we study a suboptimal but simple technique which makes it possible for each user to receive information within limited transmission periods. This scheme is later generalized to the combination of two time division multiple access (TDMA) [38,39] and opportunistic [40,41] scheduling approaches providing transmission policies for a wide range of quality-of-service requirements. Moreover, the results are compared with the ones in the standard PF scheduling scheme, which demonstrate the flexibility of the proposed methods in tackling the throughput-fairness tradeoff.

In all cases, the effect of the user outage probability on the throughput is taken into account. Particularly, the throughput is studied in the case where the user outage probability is constrained to be arbitrarily small. Finally, the simulation results and some discussions are presented in Section 5.

2 System model

2.1 Notations

In the sequel, $f_{G_n}(g_n)$ and $F_{G_n}(g_n)$ are the probability density function (pdf) and the cumulative distribution function (cdf) of the channel gain random variable G_n, respectively. Also, g_n represents the instantaneous realization of the variable G_n. A quantization encoder

$$\xi_n(g_n) = m \text{ if } g_n \in S_{n,m} = [\tilde{g}_{n,m-1}, \tilde{g}_{n,m}),\ m = 1, \ldots, M_n \tag{1}$$

is implemented at the nth receiver where M_n is the number of QRs, $S_{n,m}$ denotes the mth QR and $\tilde{g}_{n,m}$'s ($\tilde{g}_{n,0} = 0, \tilde{g}_{n,M_n} = \infty$) represent the quantization boundaries. Then, $p_{n,m} = \Pr\{g_n \in S_{n,m}\} = \int_{\tilde{g}_{n,m-1}}^{\tilde{g}_{n,m}} f_{G_n}(g_n)\mathrm{d}g_n$ is the probability that the instantaneous realization of the gain G_n falls into its mth QR.

If the nth receiver with gain $g_n \in S_{n,m}$ is scheduled, the data is sent with rate $R_{n,m} = \log(1+\hat{g}_{n,m}T_{n,m})$, where $T_{n,m}$ denotes the transmission power, and $\hat{g}_{n,m} \in S_{n,m}$ is a fixed value considered by the base station (BS). The throughput of the nth user is represented by $\bar{R}(n)$, and $\bar{R}(1, \ldots, N) = \sum_{n=1}^{N} \bar{R}(n)$ is the throughput of a system having N users.

We define the notation $u[c]\,v$ to indicate the event that the receiver u wins in the competition with v based on the selection criterion c. Moreover,

$$I^c_{u,v}(m) = \left\{j | j \in \{1, \ldots, M_v\},\ G_u \in S_{u,m},\ G_v \in S_{v,j}\,, u[c]\,v\right\} \tag{2}$$

is defined as the subset of G_v QR indices that based on the selection metric c, the receiver v loses in the competition with user u if $G_u \in S_{u,m}$. Also, $\varphi_{u,m} = \Pr\{G_v \in S_{v,j}, j \in I^c_{u,v}(m),\ \forall v \neq u\}$ denotes the probability that user u wins in the competition with the others if $G_u \in S_{u,m}$. Finally, as stated in the following, appropriate modifications are applied to the notations when semi-instantaneous fairness constraint is considered for data transmission.

2.2 Channel model

Consider the downlink of a block fading cellular system with a single BS and N receivers. While we briefly discuss random request networks, we mainly focus on fixed multiuser setups, where all users are always active, that is, N is a fixed finite number. In each time slot, a scheduler selects one of the receivers, e.g., the nth receiver. Then, the length L_c codeword $\{X[i]\,|i = 1, \ldots, L_\mathrm{c}\}$ multiplied by the fading coefficient H_n is summed with i.i.d complex Gaussian noise samples[b] $\{Z_n[i]\,|i = 1, \ldots, L_\mathrm{c}, Z_n[i] \sim \mathcal{CN}(0, \sigma_n^2)\}$ resulting in

$$Y_n[i] = H_n X[i] + Z_n[i]\,,\ i = 1, \ldots, L_\mathrm{c}. \tag{3}$$

The channel gains, defined as $G_n \doteq |H_n|^2, n = 1, \ldots, N$, are assumed to remain constant for a duration, generally determined by the channel coherence time, and then change independently according to their corresponding fading pdfs $f_{G_n}(g_n)$, $n = 1, \ldots, N$. The results are general in the sense that the channel gain distributions can be a combination of different pdfs taking positive values over the entire range $(0, \infty)$. With no loss of generality, we set the noise variances $\sigma_n^2 = 1$, $n = 1, \ldots, N$. Finally, all results are presented in natural logarithm basis, and in all simulations, the throughput is presented in nats per channel use (npcu).

Motivated by the transmission of training sequences, it is assumed that each receiver has perfect instantaneous knowledge about its own channel gain, which is an acceptable assumption for block fading networks [42-47]. On the other hand, the BS is provided with quantized CSI from all receivers. Then, as the data rate for the quantized CSI feedback is very low, it is supposed to be received noise-free and with negligible delay. This is an appropriate model for networks with stationary or slow-moving users such as wireless local area networks (WLANs). Particularly, since long block length capacity-approaching codes can be implemented in such systems, the results can provide realistic insight about the performance bounds of the considered CSI feedback approaches, e.g., [42].

2.3 Figure of merit

In delay-insensitive conditions, the ergodic capacity is a valid performance measure of fading channels [48]. Many wireless applications, however, are delay-constrained, where the codewords span a finite number of fading blocks. In this case, other performance yardsticks should be considered, among which the system throughput is the most common [43-47]. Let $R(g_1, \ldots, g_N)$ be the achievable rate, i.e., the data rate (in npcu) which can be successfully decoded by the receivers, for the gains realizations g_n, $n = 1, \ldots, N$.[c] In this way, the throughput, defined as the ratio of the expected value of decoded information nats ($E\{Q\}$) and the expected number of channel uses per block ($E\{\tau\}$) [47], is found as

$$\begin{aligned}\eta &\doteq \frac{E\{Q\}}{E\{\tau\}} = \frac{E\{L_\mathrm{c} R(g_1, \ldots, g_N)\}}{L_\mathrm{c}} \\ &= E\{R(g_1, \ldots, g_N)\} \doteq \bar{R}(1, \ldots, N),\end{aligned} \tag{4}$$

i.e., the channel average rate [43-47]. That is, with fixed-length coding, on which we focus, the throughput degen-

erates to the average rate defined as expectation on achievable rates for different gain realizations.

3 Performance analysis: unfair scenario

This section focuses on power-limited throughput optimization problem with no fairness constraint. Considering N receivers, each of which having M_n, $n = 1, \ldots, N$, CSI QRs, the quantization encoder (1) is implemented at the receivers, and the quantization indices are sent back to the BS. At the beginning of each block, the limited feedback bits are transmitted to the BS in different time/frequency subslots, such that no collision occurs in the feedback channel (a review of the different feedback transmission schemes can be found in [2]). A scheduler is employed by the BS which, based on the criterion c, selects one of the receivers to be served at any time slot[d]. Given that the user n being in region $S_{n,m}$ is selected, the data is transmitted at rate $R_{n,m} = \log(1 + \hat{g}_{n,m}T_{n,m})$, where $\hat{g}_{n,m}$ is the fixed value considered for the QR $S_{n,m}$, and $T_{n,m}$ is the considered power. If the instantaneous gain realization supports the rate, i.e., $g_n \geq \hat{g}_{n,m}$, the transmitted data is successfully decoded; otherwise outage occurs. Here, outage is defined as the event that the transmitted data is not correctly decoded by the receiver which, as the length of the codewords are asymptotically long, happens if and only if $g_n < \hat{g}_{n,m}$. Hence, for every given user n and QR $S_{n,m}$, the channel expected rate is found as

$$\begin{aligned}\bar{R}_{n,m} &= E\{\text{Achievable rates}|g_n \in S_{n,m}\}\\ &= \Pr\{\text{The } n\text{th user with } g_n \in S_{n,m} \text{ is scheduled}\}\\ &\quad \times \Pr\{\text{Successful decoding}|g_n \in S_{n,m}\}R_{n,m}\\ &= \varphi_{n,m}\Pr\left\{g_n \geq \hat{g}_{n,m}|g_n \in S_{n,m}\right\}R_{n,m}\\ &\overset{(a)}{=} \frac{\varphi_{n,m}}{p_{n,m}}\left(F_{G_n}(\tilde{g}_{n,m}) - F_{G_n}(\hat{g}_{n,m})\right)R_{n,m}.\end{aligned} \tag{5}$$

Here, (a) is based on the fact that as the nth user gain realization is in the region $S_{n,m}$, the optimal considered gain $\hat{g}_{n,m}$ must be within this region as well, i.e., $\hat{g}_{n,m} \in S_{n,m}$.

Remark 1. In practical schemes, different modulation and coding schemes are determined via the received feedback and a table of thresholds [1,2,14,19,26-28]. Here, the reconstruction points $\hat{g}_{n,m}$ work the same as the thresholds, because for a given set of powers, the transmission rate is set to $\log(1 + \hat{g}_{n,m}T_{n,m})$ if the quantization index m is received for the nth user.

The user n and the rate $R_{n,m}$ are selected by the BS if all the other gains fall into regions that lose in competition with region $S_{n,m}$. Therefore, the winning probability $\varphi_{n,m}$ is obtained by

$$\begin{aligned}\varphi_{n,m} &= \Pr\{G_\nu \in S_{\nu,j_\nu}, j_\nu \in I^c_{n,\nu}(m),\ \forall \nu = 1, \ldots, N, \nu \neq n\}\\ &= \prod_{\nu=1,\ldots,N,\nu\neq n} \Pr\{G_\nu \in S_{\nu,j_\nu}, j_\nu \in I^c_{n,\nu}(m)\}\\ &= \prod_{\nu=1,\ldots,N,\nu\neq n}\left(\sum_{j_\nu \in I^c_{n,\nu}(m)} \Pr\{G_\nu \in S_{\nu,j_\nu}\}\right)\\ &= \sum_{\substack{j_\nu \in I^c_{n,\nu}(m),\\ \nu=1,\ldots,N,\ \nu\neq n}} \cdots \sum \left(\prod p_{\nu,j_\nu}\right)\end{aligned} \tag{6}$$

which is the sum probability of all possible losing cases that may happen to the other users if $g_n \in S_{n,m}$. In this way, the throughput of the nth user is

$$\begin{aligned}\bar{R}(n) &= \sum\nolimits_{m=1}^{M_n} p_{n,m}\bar{R}_{n,m}\\ &= \sum\nolimits_{m=1}^{M_n} \varphi_{n,m}\beta_{n,m}\log\left(1 + \hat{g}_{n,m}T_{n,m}\right),\end{aligned} \tag{7}$$

where $\beta_{n,m} \doteq F_{G_n}(\tilde{g}_{n,m}) - F_{G_n}(\hat{g}_{n,m})$, and the system throughput is obtained by

$$\bar{R}(1, \ldots, N) = \sum_{n=1}^{N}\sum_{m=1}^{M_n} \varphi_{n,m}\beta_{n,m}\log\left(1 + \hat{g}_{n,m}T_{n,m}\right). \tag{8}$$

Correspondingly, the average power considered for the nth user is

$$\bar{T}(n) = \sum\nolimits_{m=1}^{M_n} p_{n,m}\varphi_{n,m}T_{n,m}, \tag{9}$$

and the system average transmission power is found as

$$\bar{T}(1, \ldots, N) = \sum\nolimits_{n=1}^{N}\sum\nolimits_{m=1}^{M_n} p_{n,m}\varphi_{n,m}T_{n,m}. \tag{10}$$

In this perspective, for a power constraint T, the power-limited throughput optimization problem can be stated as

$$\begin{aligned}&\bar{R}_{\max}(1, \ldots, N)\\ &= \max_{\forall \tilde{g}_{n,m}, \hat{g}_{n,m}, T_{n,m}} \sum\nolimits_{n=1}^{N}\sum\nolimits_{m=1}^{M_n} \varphi_{n,m}\beta_{n,m}\log\left(1+\hat{g}_{n,m}T_{n,m}\right)\\ &\qquad \text{s.t.} \quad \sum\nolimits_{n=1}^{N}\sum\nolimits_{m=1}^{M_n} p_{n,m}\varphi_{n,m}T_{n,m} \leq T\end{aligned} \tag{11}$$

which, based on the power allocation strategy and the selection criterion, can be solved numerically. Note that (8) and (10) give, respectively, the weighted sum of the users' expected achievable rate and power, where the weighting coefficients come from the scheduling properties. Also, in general, (11) is a nonconvex complex problem. Thus, it is not possible to determine the globally optimal values of all parameters analytically. However, in

the following, we present some optimality conditions for the transmission powers and the quantization parameters. Moreover, Algorithm 1, which is illustrated in Section 4, provides an iterative method for optimizing the parameters in (11).

Theorem 1. With quantized CSI and for any scheduling policy, the system throughput is bounded by

$$\bar{R}(1,\ldots,N) \geq \left(\sum_{n=1}^{N}\sum_{m=0}^{M_n}\varphi_{n,m}\beta_{n,m}\right) \times \log\left(\frac{\sum_{n=1}^{N}\sum_{m=0}^{M_n}\varphi_{n,m}\beta_{n,m}}{\sum_{n=1}^{N}\sum_{m=0}^{M_n}\frac{\varphi_{n,m}\beta_{n,m}}{\hat{g}_{n,m}T_{n,m}}}\right) \quad \text{(I)}$$

$$\bar{R}(1,\ldots,N) \leq \left(\sum_{n=1}^{N}\sum_{m=0}^{M_n}\varphi_{n,m}\beta_{n,m}\right) \times \log\left(1+\frac{\sum_{n=1}^{N}\sum_{m=0}^{M_n}\varphi_{n,m}\beta_{n,m}\hat{g}_{n,m}T_{n,m}}{\sum_{n=1}^{N}\sum_{m=0}^{M_n}\varphi_{n,m}\beta_{n,m}}\right). \quad \text{(II)} \tag{12}$$

Proof. Please see Appendix 1. □

In the following, the general problem (11) is first studied for the special case when all users have identical fading pdfs. Then, the throughput is obtained for the K-significant average feedback bit allocation scheme, and some discussions on optimal parameters in (11) are presented.

3.1 Users with the same fading pdfs and equal number of quantization regions

3.1.1 Fixed number of users

Provided that the users are homogeneous, i.e., experience the same fading pdf $f_G(g)$, and have equal number of QRs, they will use the same quantization functions

$$\xi(g) = m,\ g \in S_m = [\tilde{g}_{m-1}, \tilde{g}_m),\ m = 1,\ldots,M. \tag{13}$$

Again, M is the number of QRs, and $p_m = \int_{\tilde{g}_{m-1}}^{\tilde{g}_m} f_G(g)\mathrm{d}g$ is the probability of being in the mth region S_m. In this case, the region S_m of a user may be selected by the BS if none of the other channels fall into higher QRs. That is, for every QR S_m, we have

$$I_{n,\nu}^{c}(m) = \{j | j \leq m,\ \forall \nu = 1,\ldots,N,\ \nu \neq n\}. \tag{14}$$

Therefore, considering the nth user, the winning probability of the mth region is found as

$$\varphi_m = \sum_{\substack{\forall j_1,\ldots,j_m \in Z,\geq 0,\\ j_1+\ldots+j_m=N-1}} \cdots \sum \binom{N-1}{j_1,\ldots,j_m} \frac{p_1^{j_1}p_2^{j_2}\ldots p_m^{j_m}}{j_m+1}, \tag{15}$$

where $\binom{N-1}{j_1,\ldots,j_m} = \frac{(N-1)!}{j_1!\ldots j_m!}$, and j_k is the number of users except user n that are in the kth QR. Note that in deriving (15), we have used the fact that if along with the nth receiver, j_m of the other users fall into the mth region (and the rest are in lower QRs), one of these users is selected randomly with probability $\frac{1}{j_m+1}$. In other words, (15) gives the probability that a specific user with $g \in S_m$ is selected which occurs if none of the other user channel gains fall into the higher QRs, and the considered user is selected among the (j_m+1) users experiencing $g \in S_m$. Thus, the summation in (15) is over all possible cases, where the nth user channel gain falls in the mth QR, and the channel gains of the other users are not in the higher QRs. In this way, the throughput and the average transmission power, i.e., (8) and (10), are simplified to

$$\bar{R}(1,\ldots,N) = N\bar{R}(n) = \sum_{m=1}^{M} \left(\sum_{\substack{\forall j_1,\ldots,j_m,\\ j_1+\ldots+j_m=N-1}} \cdots \sum \frac{N!}{j_1!\ldots j_m!}\frac{p_1^{j_1}p_2^{j_2}\ldots p_m^{j_m}}{j_m+1}\right) \times \beta_m \log(1+\hat{g}_m T_m), \tag{16}$$

where $\beta_m \doteq F_G(\tilde{g}_m) - F_G(\hat{g}_m)$ and

$$\bar{T}(1,\ldots,N) = N\bar{T}(n) = \sum_{m=1}^{M} \left(\sum_{\substack{\forall j_1,\ldots,j_m,\\ j_1+\ldots+j_m=N-1}} \cdots \sum \frac{N!}{j_1!\ldots j_m!}\frac{p_1^{j_1}p_2^{j_2}\ldots p_m^{j_m+1}}{j_m+1}\right) T_m. \tag{17}$$

Here, T_m and $\hat{g}_m \in S_m = [\tilde{g}_{m-1}, \tilde{g}_m)$ are the transmission power and the fixed considered gain of each user if its gain falls into the mth QR. Replacing (16) and (17) in (11), the optimal throughput can be found based on the transmission power constraint.

Normally, there are two different interpretations of the power constraint. Due to, e.g., hardware or complexity limitations, there are cases where, independently of the feedback index, the power allocated can not exceed a maximum value T, i.e., $T_m \leq T$, $\forall m$. In this case, as the transmission rate of AWGN channels is an increasing function of the SNR [43-47], the optimal powers maximizing the throughput are obtained by $T_m = T$, $\forall m$, normally called *short-term* power allocation [43-47]. Under the more relaxed *long-term* (battery-limited) power constraint, the transmitter can adapt the power based on the channels conditions such that $\bar{T}(1,\ldots,N) \leq T$. In

this way, the optimal powers, maximizing the throughput, are found by (16), (17), and a Lagrange multiplier approach $\Upsilon = \bar{R}(1,\ldots,N) + \lambda \bar{T}(1,\ldots,N)$ leading to the water-filling equations

$$T_m = \left\lceil -\frac{\beta_m}{\lambda p_m} - \frac{1}{\hat{g}_m} \right\rceil^+. \tag{18}$$

Here, λ is the Lagrange multiplier satisfying $\bar{T}(1,\ldots,N) \leq T$ constraint and $\lceil x \rceil^+ \doteq \max(0,x)$. Intuitively, using long-term power allocation, the power is not wasted on weak channel realizations, and the saved power is spent on strong gain realizations. However, as seen in the following, the rate increment due to long-term power allocation, compared to short-term power allocation, reduces at high SNRs.

Remark 2. Equation (18) is based on the fact that replacing (16) and (17) in (11), the throughput optimization is a convex problem on power terms T_m. Therefore, (18) gives the unique optimal solution for the powers.

3.1.2 Random request network

Equation (16) demonstrates the system throughput for fixed number of receivers always requesting new information. However, there may be cases where in each time slot, only a subset of the users require information bits. We refer to this scenario as *random request network*. In this case, the inactive users send no CSI, and the scheduler selects the best candidate among the active users. Reviewing the literature, one can find different approaches modeling the users' activeness probability [22,49]. Here, we consider the model where each user becomes active with probability ϕ, independently of the others. Then, considering (16) and (17), it is easy to show that the system throughput and transmission power are found, respectively, as

$$\bar{R}^{\mathrm{RRQ}}(1,\ldots,N) = \sum_{n=1}^{N} \binom{N}{n} \phi^n (1-\phi)^{N-n} \bar{R}(1,\ldots,n) \tag{19}$$

$$\bar{T}^{\mathrm{RRQ}}(1,\ldots,N) = \sum_{n=1}^{N} \binom{N}{n} \phi^n (1-\phi)^{N-n} \bar{T}(1,\ldots,n), \tag{20}$$

where $\binom{N}{n}$ is the 'N choose n' operation. The term $\kappa_n = \binom{N}{n} \phi^n (1-\phi)^{N-n}$ in (19) to (20) gives the probability that n out of N users are active, and the summations are on all activation conditions of the N users. Then, the rate optimization problem (11) changes to

$$\begin{aligned} &\bar{R}^{\mathrm{RRQ}}_{\max}(1,\ldots,N) \\ &\quad = \max \sum_{n=1}^{N} \binom{N}{n} \phi^n (1-\phi)^{N-n} \bar{R}(1,\ldots,n) \\ &\quad \text{s.t. } \sum_{n=1}^{N} \binom{N}{n} \phi^n (1-\phi)^{N-n} \bar{T}(1,\ldots,n) \leq T \end{aligned} \tag{21}$$

which can be solved numerically.

Remark 3. Equations (19) to (21) are obtained for identically distributed users. Considering nonidentical fading channels, e.g., (19) is rephrased as

$$\bar{R}^{\mathrm{RRQ}}(1,\ldots,N) = \sum_{J \subset \{1,\ldots,N\}} \left(\left(\prod_{j \in J} \phi_j \right) \left(\prod_{j \in J^c} (1-\phi_j) \right) \times \bar{R}(\forall j \in J) \right),$$

where ϕ_j denotes the jth user activation probability, $J^c = \{1,\ldots,N\} - J$ is the complement of the set J, and $\bar{R}(\forall j \in J)$ is the average rate in the presence of the jth, $j \in J$, users which is obtained with the same procedure as in (8).

Finally, while some simulation results are presented in Section 5, we do not consider the case of users with identical pdfs in the rest of the paper.

3.2 *K*-significant average approach under short-term power constraint

Clearly, the CSI feedback overhead increases as the number of users increases. This is the point that creates challenging problems in practical systems containing large number of receivers. In order to tackle this problem, a number of suboptimal bit allocation methods have been proposed, among which we can mention the threshold-based scheduling [19,21] and *K-significant* schemes [4-7]. In the threshold-based scheduling, the users notify the transmitter only if their channel quality, e.g., SNR, exceeds some predefined threshold. However, this is a special case of the general quantized CSI feedback scheme with proper selection of the QRs.

On the other hand, under the K-significant average approach [4-7], all available feedback bits are allocated to the K most significant users, i.e., the users with the highest gain variances, and the other $(N-K)$ nonsignificant users send no feedback. With no feedback, the BS can only utilize the average characteristics of the nonsignificant users. Therefore, only one of these users, the one whose average

characteristics wins in competition with the other non-significant users, has the chance of data reception if no fairness is considered in user selection. With no loss of generality, we suppose this user to be the Nth one and let the significant users be the K first ones. The following theorem demonstrates that the throughput achieved by N users utilizing the K-significant average approach is the same as the throughput of the K significant users unless if the expected rate of the best nonsignificant user exceeds the rate obtained by the significant users all falling into their first QR.

Theorem 2. Utilizing the K-significant average approach, the optimal system throughput is obtained by

$$\begin{aligned}\bar{R}^{\text{sig}}(1,\ldots,N) = \max\Bigg\{ & \bar{R}_{\max}(1,\ldots,K), \\ & \max\{\bar{R}(1,\ldots,K) \underbrace{+p_{1,1}\times\ldots\times p_{K,1}(\vartheta-\psi)}_{(b)}\}\Bigg\},\end{aligned}$$

$$\vartheta = \left(1 - F_{G_N}(\hat{g}_N)\right)\log(1+\hat{g}_N T)$$

$$\psi = \frac{1}{p_{s,1}}\left(F_{G_s}(\tilde{g}_{s,1}) - F_{G_s}(\hat{g}_{s,1})\right)\log(1+\hat{g}_{s,1}T), \tag{22}$$

where considering uniform power allocation, $\bar{R}_{\max}(1,\ldots,K)$ is found by (11). Then, $\bar{R}(1,\ldots,K)$ is given by (8), and in each maximization, the optimization is done on all parameters $\{\tilde{g}_{n,m}, \hat{g}_{n,m}, n = 1,\ldots,K, m = 1,\ldots,M_n, \hat{g}_N\}$. Also, s is the best user selected among significant users if all fall in their first QR, and ϑ is the expected rate of the Nth user with no CSI feedback. Then, the term (b) is the rate increase/decrease obtained by communicating the non-significant Nth user if it is selected by the scheduler when the significant users fall into their first QRs.

Proof. Comparing the Nth and any of the K significant users, it is obvious that this is not optimal to consider more than one QR for the significant users that lose in competition with the Nth user, as the losing QRs can be merged together. That is, the Nth receiver can be served by the BS only when all significant users fall into their first QR and the Nth user wins in competition with them. In this way, the optimal system throughput is obtained by (22). Here, (22) checks whether scheduling the nonsignificant user increases the throughput or the significant users should always be scheduled even if all of them fall into their first QRs. □

With the same argument, the results can be extended to the case when optimal power allocation is implemented by the BS. Also, although suboptimal, K-significant average feedback bit allocation approach is a simple procedure with no additional bit mapping requirement, which makes it interesting in practical CSI feedback schemes dealing with nonidentical gain pdfs [4-8]. Note that along with the K-significant average, there is a K-significant *instantaneous* scheme where, in each block, the feedback bits are dynamically allocated to the users experiencing the highest channel quality. However, the K-significant instantaneous method, which is normally used in OFDMA systems, is not implementable in our communication setup as the users are not connected and do not have access to the instantaneous CSI of each other. Finally, the simulation results of the K-significant average approach can be found in Section 5.

3.3 Some discussions on the optimality conditions of the power-limited throughput maximization problem

This part presents some discussions on the optimality conditions of the quantization parameters involved in (11).

Theorem 3. Under short-term power constraint $T_{n,m} = T, \forall n,m$, the optimal quantization parameters satisfy the condition $\tilde{g}_{n,m} = \hat{g}_{n,m+1}, \forall n,m$, independent of the scheduling policy, channels distributions, and the number of quantization regions or the users.

Proof. Please see Appendix 2. □

Theorem 3 studied the optimal quantization parameters under the short-term power constraint. In order to reprove the same optimality conditions under the premise of optimal power allocation, we should first show that to maximize the system throughput, the higher quantization regions should receive more power, i.e., $T_{n,m} \leq T_{n,m+1}, \forall n,m$, if $\varphi_{n,m} \leq \varphi_{n,m+1}$. For this reason, two adjacent QRs of a user are considered, and it is shown that the higher region has more contribution on the system throughput and so should receive more power. Based on (5), the contributions of the mth and the $(m+1)$th QRs of the nth user on the system throughput are $\bar{R}_{n,m} = \varphi_{n,m}\Pr\{g_n \geq \hat{g}_{n,m} | g_n \in S_{n,m}\}\log(1+\hat{g}_{n,m}T_{n,m})$ and $\bar{R}_{n,m+1} = \varphi_{n,m+1}\Pr\{g_n \geq \hat{g}_{n,m+1} | g_n \in S_{n,m+1}\}\log(1+\hat{g}_{n,m+1}T_{n,m+1})$, respectively. Also, the average power consumed in these two regions is

$$\chi = \varphi_{n,m}p_{n,m}T_{n,m} + \varphi_{n,m+1}p_{n,m+1}T_{n,m+1}.$$

Setting $\hat{g}_{n,m+1} = \hat{g}_{n,m} \in S_{n,m}$, which is obviously nonoptimal for $\hat{g}_{n,m+1}$, we have $\Pr\{g_n \geq \hat{g}_{n,m+1} | g_n \in S_{n,m+1}\} = 1$

which leads to $\bar{R}_{n,m+1}^{\text{nonoptimal}} = \varphi_{n,m+1}\log(1+\hat{g}_{n,m}T_{n,m+1}) \leq \bar{R}_{n,m+1}$. Then, comparing $\bar{R}_{n,m+1}^{\text{nonoptimal}}$ and $\bar{R}_{n,m}$, it is obvious that in the optimal case, in terms of (11), we have

$$T_{n,m} \leq T_{n,m+1}, \forall n, m.$$

This is particularly due to the fact that by changing the set of powers $(T_{n,m}, T_{n,m+1})$, where $T_{n,m} = T_{n,m+1}$, to $(T_{n,m} - \varepsilon, T_{n,m+1} + \frac{\varphi_{n,m}p_{n,m}}{\varphi_{n,m+1}p_{n,m+1}}\varepsilon)$, $\varepsilon \geq 0$, the sum $\bar{R}_{n,m+1}^{\text{nonoptimal}} + \bar{R}_{n,m}$ (and consequently $\bar{R}_{n,m+1} + \bar{R}_{n,m}$) is increased while the average power consumption χ is kept the same as before. Therefore, in order to have maximum throughput, the power should be preferably given to the last QRs.

Note that the property $\varphi_{n,m} \leq \varphi_{n,m+1}$, i.e., $I_{n,u}^{c}(m) \subset I_{n,u}^{c}(m+1)$, simply means that the scheduler should be designed such that the better the channel quality of a user is, the higher winning chance it has. To the best of the authors' knowledge, this is one of the properties of all the present scheduling schemes.

In this way, the same argument as for the short-term power constraint is used to show the validity of Theorem 3 under long-term power constraint; assume that the quantization parameters and the powers have been optimized such that (8) is maximized but $\tilde{g}_{n,m} \neq \hat{g}_{n,m+1}$. If the current gain realization of the winner user n is within the region $\tilde{g}_{n,m} \leq g_n < \hat{g}_{n,m+1}$, the data transmitted at rate $\log(1+\hat{g}_{n,m+1}T_{n,m+1})$ is lost, and the power $T_{n,m+1}$ is wasted. On the other hand, by reducing the rate to $\log(1+\hat{g}_{n,m}T_{n,m})$ not only the system throughput is increased by $\varphi_{n,m}(F_{G_n}(\hat{g}_{n,m+1}) - F_{G_n}(\tilde{g}_{n,m}))\log(1+\hat{g}_{n,m}T_{n,m})$ but also the average transmission power is reduced, as $T_{n,m} \leq T_{n,m+1}$. The rate increment and power reduction conflict the optimality assumption, and so in the optimal case, we have $\tilde{g}_{n,m} = \hat{g}_{n,m+1}, \forall n, m$.

Figure 1 demonstrates the quantizers' optimality condition more clearly. This is an intuitive result, meaning that to maximize the throughput in power-limited communication setups utilizing scheduling and quantized CSI feedback, the channel gain should be assumed to be constant, equal to its worst case within each QR except the first one. Finally, note that the arguments are valid for both fixed and random request networks.

According to the above discussions, optimizing the power-limited system throughput, the outage *may* happen iff a user in the first QR is selected; in the optimal case, the codeword transmitted to the scheduled user within $m > 1$ QR is always decoded, as the data is sent with the lowest possible rate. Therefore, the outage only happens if (1) a user in the first QR is scheduled and (2) $g_n < \hat{g}_{n,1}$ where n is the index of the scheduled user (the outage region in Figure 1b).

Interestingly, the system outage probability vanishes as the number of users goes to infinity. This is because with infinitely many users, i.e., $N \to \infty$, the probability $\Pr\{\text{User } n \text{ is scheduled } \& \, g_n < \hat{g}_{n,1}\}$ goes to zero, as there are always users with high channel quality. That is, for every given threshold ζ, we have $\lim_{N\to\infty} \Pr\{\forall \zeta, \exists k \in \{1,\ldots,N\}, g_k \geq \zeta\} \to 1$. Hence, no outage occurs as the users with high quantization indices are always scheduled.

Corollary 1. For every given long-term power constraint, we have $T_{n,m} > 0$, $m > 1$ as long as $\varphi_{n,m} \leq \varphi_{n,m+1}$. That is, the only region that *may* receive no transmission power is the first one.

Proof. The property $T_{n,m} \leq T_{n,m+1}, \forall n, m$, simply means that for each user under long-term power allocation strategy, no power is assigned to a QR until the higher regions have received their required power. That is, for each power constraint T, there is an index $\breve{m}$ of the QRs, where $T_{n,m} > 0$ for $m \geq \breve{m}$ and $T_{n,m} = 0$ for $m < \breve{m}$. However, in the optimal case, the QRs $1 \leq m < \breve{m}$ that have received no power are merged together to make a single (first) QR. Hence, $T_{n,m} > 0$ if $m > 1$. This is a useful conclusion simplifying the optimal power allocation algorithm. □

Finally, we close the section with the following discussion about the transmission parameters setting. In practice, the suitable transmission parameters can be determined in two ways. In the first method, the system performance is evaluated off-line for different rates/powers, and the appropriate parameters are collected in a table which is used during data transmission. In this case, which is the same as in adaptive modulation and coding protocols [1], there is no need to know the channel cdf (in general, the only parameters that we need to know are the QRs and the transmission powers, and not the channel cdf). In the second method, however, the gain cdfs and an optimization algorithm are utilized by the BS for parameter setting. This is a suitable method for the scenario where the channel follows a specific cdf pattern, and only the average statistics (e.g., the gains mean and variance) change slowly. Since the average statistics of the channel vary slowly (in most cases with a time constant of several 100 ms), the amount of feedback needed for long-run adaptation will be negligible compared to the channel quantization feedback.

4 Performance analysis: fairness scenario

In a homogeneous system, with statistically identical users, scheduling the user with the strongest channel at any given time slot maximizes not only the total system

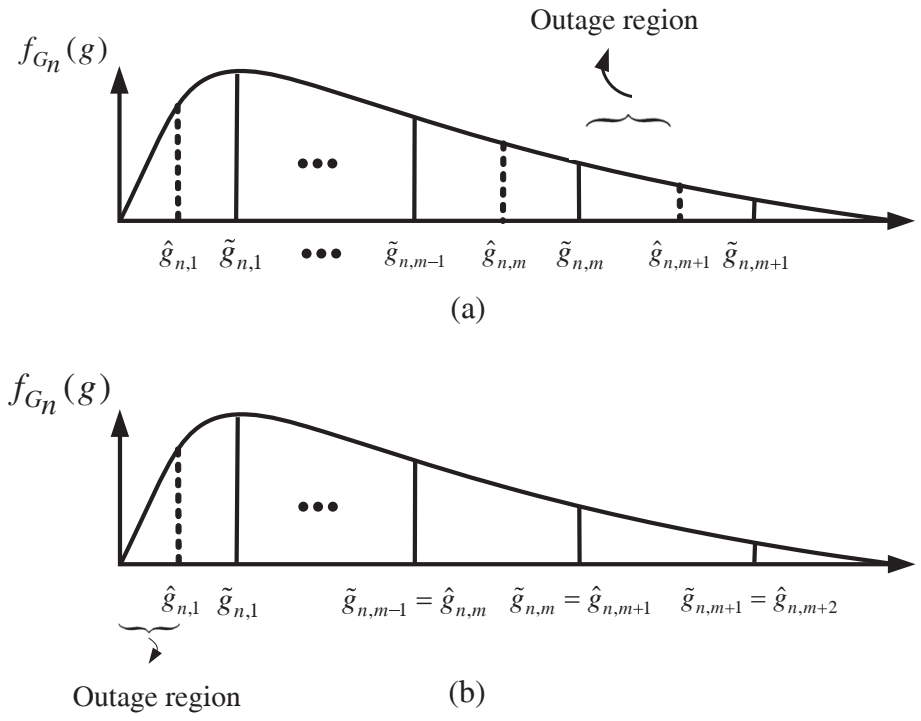

Figure 1 Quantization boundaries of a channel fading pdf. (a) Nonoptimal, **(b)** Optimal.

throughput but also the long-term data transmission efficiency of individual users. However, in a typical wireless network, the channel distributions are not necessarily identical. In this case, the strategy of communicating with the user having the best channel leads to giving almost all the resources to the statistically stronger user(s), thereby resulting in an unfair scheduling. A scheduling algorithm with fairness should sacrifice part of the system throughput and multiuser diversity to equalize the probability that a user is scheduled. There are many schemes tackling the fairness problem [15-17,34-37]. Here, we focus on two different fairness techniques as follows.

4.1 Long-term fairness constraint

There may be applications where while optimizing the system power-limited data transmission efficiency, every individual user should be guaranteed to have a given minimum throughput in long term. We call this limitation as long-term fairness constraint where, using (7), the rate optimization problem (11) is replaced by

$$\bar{R}^{\mathrm{LT}}_{\max}(1,\ldots,N) = \max_{\forall \tilde{g}_{n,m},\hat{g}_{n,m},T_{n,m}} \sum_{n=1}^{N}\sum_{m=1}^{M_n} \varphi_{n,m}\beta_{n,m}\log(1+\hat{g}_{n,m}T_{n,m})$$
$$\text{s.t.} \begin{cases} \sum_{n=1}^{N}\sum_{m=1}^{M_n} p_{n,m}\varphi_{n,m}T_{n,m} \le T, & \text{(I)} \\ \sum_{m=1}^{M_n} \varphi_{n,m}\beta_{n,m}\log(1+\hat{g}_{n,m}T_{n,m}) \ge \gamma_n, \forall n & \text{(II)} \end{cases} \tag{23}$$

Here, γ_n is the nth user throughput constraint. Also, (23.II) constrains each user to have a given minimum average rate. That is, the power-limited throughput is maximized under the condition that the user individual throughput constraints $\bar{R}(n) \ge \gamma_n, \forall n$, are satisfied. Note that (23) is changed to (11) if (23.II) is relaxed. Finally, with the same procedure as before, we have $\hat{g}_{n,m+1} = \tilde{g}_{n,m}$, $m \ge 1$, in the optimal case.

Depending on the user selection criterion and the power allocation strategy, there may be no closed-form solution for (23). This is particularly because the best user, in the sense of *unfair* throughput, i.e., (11), should not necessarily be selected, to satisfy (23.II) for all users. In simple words, in order to get some constrained system throughput, we may need to devote some rates for nonoptimal, in terms of throughput, receivers such that (23.II) is satisfied[e]. Therefore, (23) leads to a complex nonconvex optimization problem even in its simplest cases. To tackle this problem under short-term power constraint, we propose a numerical method stated in Algorithm 1 which solves the problem using optimal scheduling, in the sense of (23).

Algorithm 1 Throughput optimization

I. For a given short-term power allocation constraint $T_{n,m} = T$, $\forall n, m$, and the user individual throughput constraints $\gamma_n, n = 1 \dots N$, consider J, e.g., $J = 20$, randomly generated vectors $\Omega^{(j)} = [\tilde{g}_{1,1}^{(j)}, \dots \tilde{g}_{N,M_N}^{(j)}, \hat{g}_{1,1}^{(j)}, \dots, \hat{g}_{N,1}^{(j)}]$, where $0 < \hat{g}_{n,1}^{(j)} \leq \tilde{g}_{n,1}^{(j)} \leq \tilde{g}_{n,2}^{(j)} < \dots < \tilde{g}_{n,M_n}^{(j)}, \forall n$.

II. For each vector $\Omega^{(j)}$, do the following subprocedures:

   1. For all possible user selection policies, i.e., all possible QR selection configurations, find the throughput based on (8).
   2. For each user selection configuration, set the throughput equal to zero, if (23.II) is not satisfied.
   3. Set $\bar{R}^{(j)}(1, \dots, N)$, i.e., the throughput of the jth vector, equal to the maximum of the throughput obtained under different user selection configurations.

III. Determine the vector which results in highest throughput, i.e., $\Omega^{(k)}$, where $\bar{R}^{(j)}(1, \dots, N) \leq \bar{R}^{(k)}(1, \dots, N)$, $\forall j = 1, \dots, J$.

IV. $\Omega^{(1)} \leftarrow \Omega^{(k)}$.

V. Generate $q \ll J$, e.g., $q = 5$, vectors $\Omega^{(i),\text{new}}$, $i = 1, \dots, q$ around $\Omega^{(1)}$. These vectors are generated by adding small random numbers to the coefficients of $\Omega^{(1)}$ such that they satisfy the constraints introduced in I.

VI. $\Omega^{(i+1)} \leftarrow \Omega^{(i),\text{new}}$, $i = 1, \dots, q$.

VII. The same as in step I, regenerate the remaining vectors $\Omega^{(j)}, j = q + 2, \dots, J$ randomly.

VIII. Go to II and continue until convergence.

Algorithm 1 roots from the genetic algorithm concepts which, the same as other machine learning-based algorithms such as particle swarm optimization, is based on (1) searching around the current solution, i.e., testing solutions with small random changes in the parameters of the current solution as well as (2) reducing the effect of local minima by mutation. In each iteration of the algorithm, we search around the best solution which has been obtained up to now and check a number of randomly generated answers which help prevent it from getting trapped into the local minima.

Similar to other techniques for solving nonconvex optimization problems, it can not be guaranteed that the algorithm leads to the globally optimal solution for all channel conditions. However, the algorithm can be run for many different initial parameter settings to reduce the effect of local minima. Furthermore, our experiments show that the algorithm is much more efficient than using a greedy search scheme which requires a large number of initial random seeds due to the nonconvexity of (23). Finally, although it may be time consuming when the number of optimization parameters increases, the proposed algorithm has been shown to be efficient in many complex optimization problems dealing with local minima issues [50].

4.2 Semi-instantaneous fairness constraint

The optimization problem (11) determines the maximum long-term system performance in power-limited conditions. In this case, depending on the gains realizations, it may happen that a user remains off for a long time. On the other hand, in delay-sensitive applications, the users may require to be served not always but repeatedly within limited transmission periods. In other words, the scheduling approach should guarantee that each user gets some minimum amount of information in limited periods, to keep the connection alive. In order to provide such a property, we consider a simple scheme where, deviating from the optimal scheduling approach, the receivers are served by the standard round robin (RR) approach for a certain amount of time. In this way, while optimal scheduling, in the sense of (11), is utilized in most of fading blocks, in some time slots, the scheduler is switched off and the users are communicated one by one.

Setting $\varphi_{n,m} = 1, \forall m$, the throughput and the average transmission power of the single-input single-output (SISO) system between the transmitter and the nth receiver, with no competition with others, are obtained by

$$\bar{\theta}(n) = \sum_{m=1}^{M_n^{\text{r}}} \left(F_{G_n}(\tilde{g}_{n,m}^{\text{r}}) - F_{G_n}(\hat{g}_{n,m}^{\text{r}})\right) \log(1 + \hat{g}_{n,m}^{\text{r}} T_{n,m}^{\text{r}}) \tag{24}$$

and

$$\bar{T}_n^{\text{r}} = \sum_{m=1}^{M_n^{\text{r}}} p_{n,m}^{\text{r}} T_{n,m}^{\text{r}}, \tag{25}$$

respectively. Here, M_n^{r} is the number of QRs, $p_{n,m}^{\text{r}} = \int_{\tilde{g}_{n,m-1}^{\text{r}}}^{\tilde{g}_{n,m}^{\text{r}}} f_{G_n}(g_n) \text{d}g_n$, and $\tilde{g}_{n,m}^{\text{r}}$ represents the mth quantization boundary of the encoder function

$$\xi_n^{\text{r}}(g_n) = m \text{ if } g_n \in S_{n,m}^{\text{r}} = [\tilde{g}_{n,m-1}^{\text{r}}, \tilde{g}_{n,m}^{\text{r}}),\ m = 1, \dots, M_n^{\text{r}}$$

which is implemented within the RR-based transmission periods of the nth user. Also, $T_{n,m}^{\text{r}}$ and $\hat{g}_{n,m}^{\text{r}} \in [\tilde{g}_{n,m-1}^{\text{r}}, \tilde{g}_{n,m}^{\text{r}})$ denote, respectively, the transmission power and the fixed considered gain if the nth channel falls in the mth QR

$S^{\mathrm{r}}_{n,m}$. In this way, considering the RR period to be α, the constrained throughput is found as

$$\bar{R}^{\mathrm{semi}}(1,\ldots,N)=(1-\alpha)\bar{R}(1,\ldots,N)+\frac{\alpha}{N}\sum_{n=1}^{N}\sum_{m=1}^{M^{\mathrm{r}}_n}\left(F_{G_n}(\tilde{g}^{\mathrm{r}}_{n,m})-F_{G_n}(\hat{g}^{\mathrm{r}}_{n,m})\right)\log(1+\hat{g}^{\mathrm{r}}_{n,m}T^{\mathrm{r}}_{n,m}), \tag{26}$$

and the average transmission power of the system is

$$\bar{T}^{\mathrm{semi}}(1,\ldots,N)=(1-\alpha)\bar{T}(1,\ldots,N)+\frac{\alpha}{N}\sum_{n=1}^{N}\sum_{m=1}^{M^{\mathrm{r}}_n}p_{n,m}T^{\mathrm{r}}_{n,m}, \tag{27}$$

where $\bar{R}(1,\ldots,N)$ and $\bar{T}(1,\ldots,N)$ are obtained by (8) and (10), respectively. The first term in (26) represents the throughput achieved during the scheduler activation period, and the second term gives the throughput achieved during the RR period. Consequently, the power-limited rate optimization problem, i.e., (11), is rephrased as

$$\begin{aligned}&\bar{R}^{\mathrm{semi}}_{\max}(1,\ldots,N)\\&\quad=\max\left\{(1-\alpha)\bar{R}(1,\ldots,N)+\frac{\alpha}{N}\sum\nolimits_{n=1}^{N}\sum\nolimits_{m=1}^{M^{\mathrm{r}}_n}\left(F_{G_n}(\tilde{g}^{\mathrm{r}}_{n,m})-F_{G_n}(\hat{g}^{\mathrm{r}}_{n,m})\right)\log(1+\hat{g}^{\mathrm{r}}_{n,m}T^{\mathrm{r}}_{n,m})\right\}\\&\quad\text{s.t. }(1-\alpha)\bar{T}(1,\ldots,N)+\frac{\alpha}{N}\sum\nolimits_{n=1}^{N}\sum\nolimits_{m=1}^{M^{\mathrm{r}}_n}p^{\mathrm{r}}_{n,m}T^{\mathrm{r}}_{n,m}\le T.\end{aligned} \tag{28}$$

With the same arguments as before, we have $\hat{g}^{\mathrm{r}}_{n,m+1}=\tilde{g}^{\mathrm{r}}_{n,m}$, $\hat{g}_{n,m+1}=\tilde{g}_{n,m}$, $m\ge 1$ in this case as well. Finally, implementing appropriate modifications in Algorithm 1, the optimal parameters $\hat{g}_{n,1},\tilde{g}_{n,m},T_{n,m},\hat{g}^{\mathrm{r}}_{n,1},\tilde{g}^{\mathrm{r}}_{n,m},T^{\mathrm{r}}_{n,m}$, $\forall n,m$ can be determined numerically. For this case, we set $\gamma_n=0,\forall n$, in the algorithm and replace $\Omega^{(j)}$ and $\bar{R}^{(j)}(1,\ldots,N)$ with $\Omega^{(j)}=[\tilde{g}^{(j)}_{1,1},\ldots\tilde{g}^{(j)}_{N,M_N},\hat{g}^{(j)}_{1,1},\ldots,\hat{g}^{(j)}_{N,1},\tilde{g}^{\mathrm{r},(j)}_{1,1},\ldots\tilde{g}^{\mathrm{r},(j)}_{N,M^{\mathrm{r}}_N},\hat{g}^{\mathrm{r},(j)}_{1,1},\ldots,\hat{g}^{\mathrm{r},(j)}_{N,1}]$ and $\bar{R}^{\mathrm{semi},(j)}(1,\ldots,N)$, respectively, where the function $\bar{R}^{\mathrm{semi},(j)}(1,\ldots,N)$ is found according to (26).

Although the proposed scheme gives the users the chance of data reception in limited time slots, there is still a positive probability that they receive no information even within the round robin periods; using the worst case condition, i.e., $\hat{g}^{\mathrm{r}}_{n,m+1}=\tilde{g}^{\mathrm{r}}_{n,m}$, the transmitted data of the nth user is lost if its instantaneous gain realization g_n falls in the region $g_n\in[0,\hat{g}^{\mathrm{r}}_{n,1})$ (Corollary 1). Hence, with limited number of QRs, there is always a positive outage probability $\pi^{\mathrm{r}}_n=\Pr\{0\le g_n<\hat{g}^{\mathrm{r}}_{n,1}\}=F_{G_n}(\hat{g}^{\mathrm{r}}_{n,1})$ for each user during the round robin time slots. However, by adding the $\hat{g}^{\mathrm{r}}_{n,1}\le F^{-1}_{G_n}(\xi_n)$ constraint to the optimization problem, i.e., (28), this outage probability of the users can be reduced to a given threshold ξ_n. In this way, we guarantee that with probability of $(1-\xi_n)$, the user n receives some information within a limited number of transmission blocks determined by α.

4.2.1 Generalization of the fairness scheme

Equation (28) is derived based on the assumption that the nonoptimally scheduled, in the sense of (11), part of the transmission periods is uniformly allocated to the users (round robin scheme). However, in order to provide a better fairness scheme, these time slots can be divided between the receivers in a wiser manner such that the users' quality-of-service requirements are satisfied. In this perspective, the constrained throughput optimization problem (28) is generalized to

$$\begin{aligned}&\bar{R}^{\mathrm{semi}}_{\max}(1,\ldots,N)\\&\quad=\max\left\{(1-\alpha)\bar{R}(1,\ldots,N)+\sum\nolimits_{n=1}^{N}\sum\nolimits_{m=1}^{M^{\mathrm{r}}_n}\alpha_n\left(F_{G_n}(\tilde{g}^{\mathrm{r}}_{n,m})-F_{G_n}(\hat{g}^{\mathrm{r}}_{n,m})\right)\log(1+\hat{g}^{\mathrm{r}}_{n,m}T^{\mathrm{r}}_{n,m})\right\}\\&\quad\text{s.t.}\begin{cases}(1-\alpha)\bar{T}(1,\ldots,N)+\sum_{n=1}^{N}\sum_{m=1}^{M^{\mathrm{r}}_n}\alpha_n p^{\mathrm{r}}_{n,m}T^{\mathrm{r}}_{n,m}\le T,\\ \hat{g}^{\mathrm{r}}_{n,1}\le F^{-1}_{G_n}(\xi_n)\ \forall n,\\ \sum_{n=1}^{N}\alpha_n=\alpha\le 1,\ \alpha_n\ge 0,\end{cases}\end{aligned} \tag{29}$$

where α_n is the fraction of the time slots in which the user n is selected regardless of the other channels' quality. Here, the constraint $\hat{g}^{\mathrm{r}}_{n,1}\le F^{-1}_{G_n}(\xi_n)$ limits the nth user outage probability during the fairness-based data transmission period to a given threshold ξ_n. Again, we have $\tilde{g}_{n,m}=\hat{g}_{n,m+1}$, $\tilde{g}^{\mathrm{r}}_{n,m}=\hat{g}^{\mathrm{r}}_{n,m+1}$, $\forall n,m\ge 1$ in the optimal case.

Remark 4. Equation (29) is a generalization of the two TDMA-based ($\alpha=1$) [13,38,39] and opportunistic ($\alpha=0$) [40,41] scheduling schemes which can determine the optimal data transmission strategy in a wide range of communication scenarios.

Theorem 4. Let $\bar{R}^{\mathrm{semi}}_{\max}(1,\ldots,N|M^{\mathrm{r}}_n,\xi_n)$ be the maximum throughput obtained by (29) when the nth user has an outage probability constraint ξ_n and M^{r}_n QRs in the RR-based time slots. Then, under a short-term power constraint, $\lim_{\xi_n\to 0}\bar{R}^{\mathrm{semi}}_{\max}(1,\ldots,N|M^{\mathrm{r}}_n,\xi_n)=\bar{R}^{\mathrm{semi}}_{\max}(1,\ldots,N|M^{\mathrm{r}}_n-1,1)$. That is, the RR-based throughput when the outage probability of a user tends to zero is equal to the throughput

with relaxed outage probability constraint and one less region in the quantizer of that user.

Proof. Please see Appendix 3. □

In other words, the theorem simply means that in order to have an outage-free data transmission within the RR-based time slots, the data should be transmitted with an extremely low rate when the channel falls in the first QR. Therefore, the first QR has (almost) no contribution on the system throughput which will be determined by averaging on the achievable rates of the $m \geq 2$ QRs.

Remark 5. Using Theorem 3 and Figure 1b, the same assertion can be proved for the case when an outage probability constraint is considered in the opportunistic scheduling period. Note that in both fairness schemes considered in this paper, the parameters are determined off-line, and consequently, fairness does not increase the implementation complexity or the feedback load. Moreover, in both methods, the users' contributions on the throughput are not necessarily the same, and they can be *weighted*. This gives flexibility, for instance compared to the standard PF scheduling, in dealing with the throughput-fairness tradeoff.

5 Simulation results and discussions

Simulation results are achieved for Rayleigh fading channels $f_{G_n}(g_n) = \frac{1}{\mu_n} e^{-\frac{g_n}{\mu_n}}, g_n > 0$, where μ_n denotes the nth user fading parameter. Also, the noise variances are set to $\sigma_n^2 = 1$, $n = 1, \ldots, N$. Figures 2, 3, 4, 5, and 6 concentrate on the case where the users experience identical fading distributions with $\mu_n = 1, \forall n$. Here, the optimal quantizers have been obtained with proper modifications in Algorithm 1. Also, the results have been double-checked by the modified version of the gradient-based quantization algorithm of [43] which studies the SISO channels. The results are presented as follows.

5.1 System throughput with different power constraints and user activation probabilities

Figure 2 demonstrates the system throughput for the fixed network scenario and under different transmission power constraints. Then, the system performance in random request networks is studied in Figures 3 and 4, where the throughput is obtained under different transmission power constraints and user activation probabilities. Here, the results show that

- Substantial throughput increment is achieved via very limited number of QRs, i.e., feedback bits, (Figures 2, 3, and 4). Particularly, the increment is considerable when the user activation probability increases (Figure 3). Also, in harmony with the literature, the partial CSI is more effective when the number of users increases. The intuition behind this is that with more CSI, the multiuser diversity which is an increasing function of the number of users is more efficiently exploited.

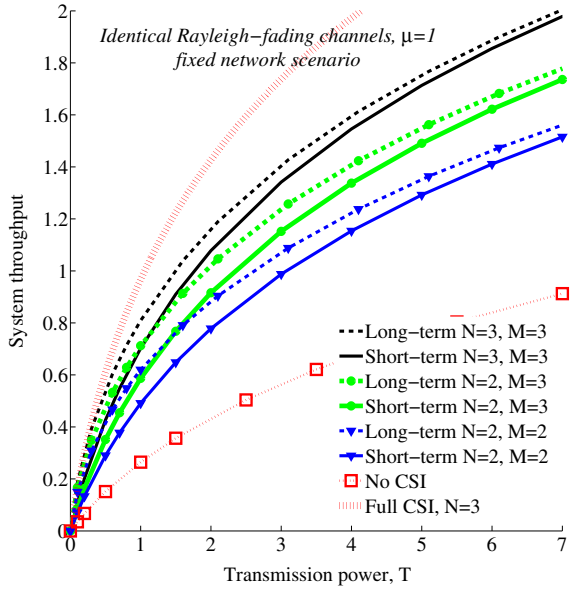

Figure 2 Throughput versus transmission power, *T*. Identical fading pdfs, $\mu = 1$, fixed network scenario. *N* is the number of users, and *M* represents the number of QRs considered for each user.

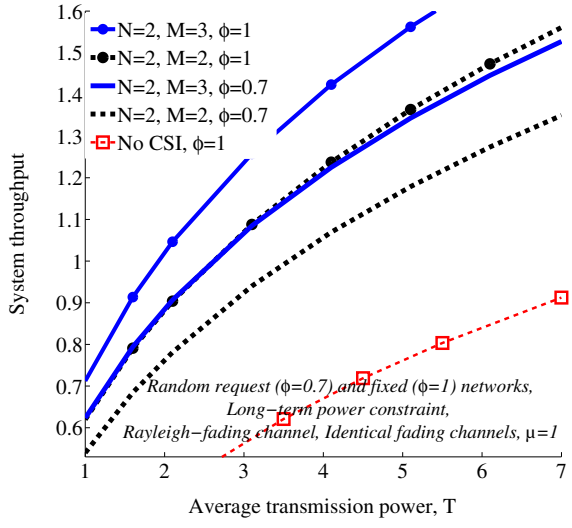

Figure 3 Throughput versus transmission power constraint *T* and different user activation probabilities. Long-term power constraint, identical fading pdfs, $\mu = 1$.

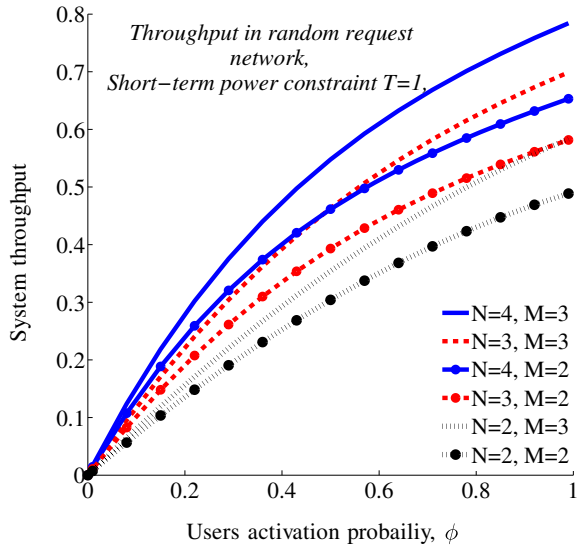

Figure 4 Throughput versus user activation probability ϕ. Short-term power constraint $T = 1$, identical fading pdfs, $\mu = 1$.

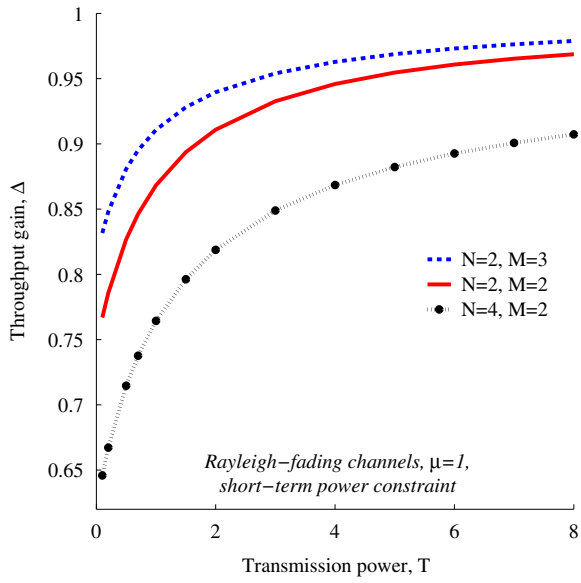

Figure 6 Throughput gain versus the transmission power *T*, identical fading pdfs, $\mu = 1$, short-term power constraint.

- The relative effect of optimal power allocation, compared to short-term power allocation, is slightly higher when the number of users decreases (Figure 2).

5.2 The optimal quantization parameters

Considering $M = 3$ QRs for each user, Figure 5 shows the optimal reconstruction points $\hat{g}_m$ obtained for different transmission powers and number of users. Here, the following points are interesting to be noted:

- According to Theorem 3 and its following discussions, the last two points in each triple of points $(\hat{g}_1, \hat{g}_2, \hat{g}_3)$ coincide with the quantization boundaries, while the first point is the fixed value $\hat{g}_1$ considered in the first region. Also, the same points are obtained by $M = 4$ QRs and extremely hard outage probability constraint for each user, where the first QR converges to zero (Remark 5).
- The optimal QRs get closer to zero when the transmission power increases.
- The QRs expand as the number of users increases. The intuition behind this point is that with high number of users, the probability that a user with low channel quality is scheduled reduces. Therefore, the quantization boundaries expand such that high SNRs are quantized with higher resolution.

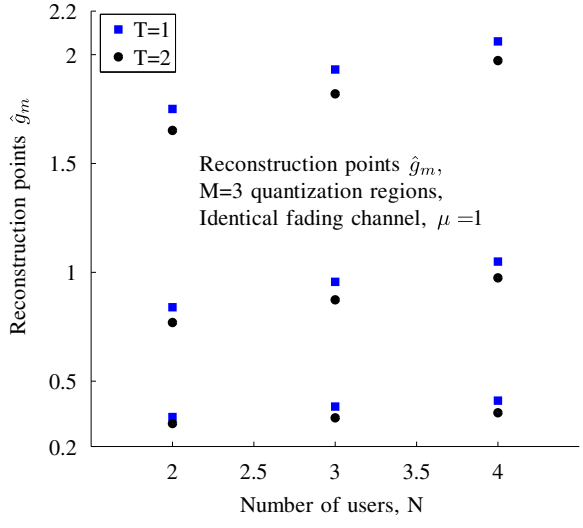

Figure 5 Optimal reconstruction points for different number of users and transmission powers. Identical fading pdfs, $\mu = 1$, $M = 3$ quantization regions.

5.3 The system throughput in the presence of different quantizers

The effect of optimal channel quantization on the system throughput is investigated in Figure 6. Here, the throughput gain $\Delta = \frac{\bar{R}^{\text{Eq}}}{\bar{R}^{\text{Opt}}}$ is plotted as a function of the transmission power, where $\bar{R}^{\text{Opt}}$ and $\bar{R}^{\text{Eq}}$ are the system throughput obtained by optimal channel quantization and equal-probability quantizer, respectively (with equal-probability quantization, the channel pdf is divided into M regions having the probabilities $\frac{1}{M}$). According to the figure, the following points are concluded:

- The channel quantization has a large impact on the system performance, particularly at low SNRs. For instance, with $N = 4$ users and one bit feedback per user, i.e., $M = 2$, about 35% of the throughput is lost

if the optimal quantizers are replaced with equal-probability quantization at low SNRs.
- As expected, optimal quantization becomes more important when the number of users increases.
- On the other hand, the performance loss due to nonoptimal quantizers decreases when the number of QRs or the transmission power increases.

Figures 7, 8, 9, and 10 focus on the system throughput and fairness approaches in the presence of nonidentical fading pdfs.

5.4 Throughput in the presence of *K*-significant average feedback bit allocation approach

Assuming $K = 2$- and $K = 3$-significant approach and $T = 1$, Figure 7 presents the system throughput as a function of a nonsignificant user fading parameter. Here, the K significant users are supposed to have identical fading coefficients $\mu = 1$, while the nonsignificant user fading coefficient is less than one. Note that the number of users can be any number $N \geq K + 1$, as only the best nonsignificant user participates in the scheduling. The figure emphasizes the following points:

- Using K-significant average approach, the nonsignificant users affect the system throughput only when their average characteristics are comparable with ones in the significant users; when the nonsignificant users experience very poor channel condition, they have no contribution to the system throughput, since the significant users are always scheduled (flat lines in Figure 7). However, when the nonsignificant users' channel quality improves, they can win in competition with the significant users falling into their first QRs (the increasing parts of the curves in Figure 7; also, please see Theorem 2).
- The nonsignificant users have less chance of data transmission, i.e., the increasing parts start at higher values of μ, when the number of users increases. This is intuitively due to the fact that with higher number of significant users, the probability that all users experience bad channel condition, i.e., all fall into first QR, reduces.
- On the other hand, the nonsignificant users are more involved in the scheduling process when the number of QRs considered for the significant users increases. This is because with more QRs, the first region and its contribution to the throughput become smaller.

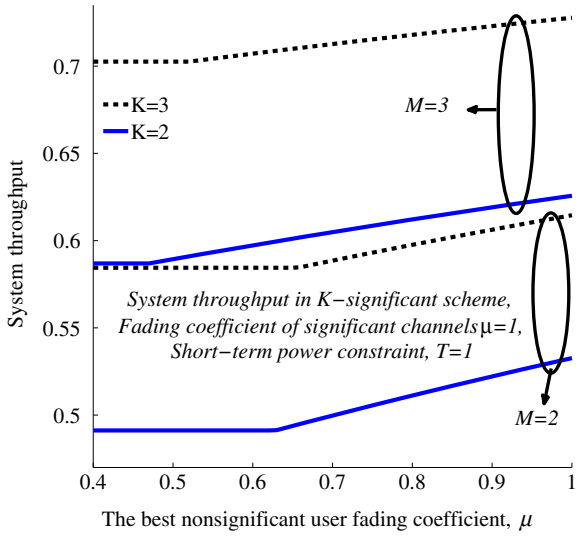

Figure 7 Throughput versus the nonsignificant user fading parameter. K-significant average approach, short-term power constraint $T = 1$, fading coefficient of significant users $\mu = 1$. The total number of users can be $N \geq K + 1$ as long as the fading coefficient of the other nonsignificant users is less than μ.

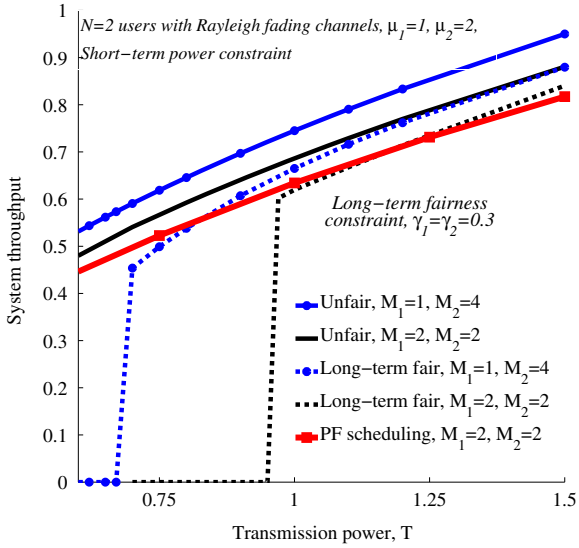

Figure 8 The throughput versus the transmission power *T* under long-term fairness constraint, $\gamma_1 = \gamma_2 = 0.3$. Short-term power constraint, $N = 2$ users with fading coefficients $\mu_1 = 1, \mu_2 = 2$. With a long-term fairness constraint, no transmission is allowed at low SNRs, as the users' individual throughput constraints (23.II) are not satisfied. Increasing the power, there is a threshold above which data transmission is possible since the users' individual throughput constraints are satisfied.

5.5 The long-term fairness constraint and feedback bit allocation strategies

Figure 8 demonstrates the system throughput under long-term fairness constraint. Here, two nonidentical users are considered where the users' fading coefficients and the throughput thresholds are set to $\mu_1 = 1, \mu_2 = 2$, and $\gamma_1 = \gamma_2 = 0.3$, respectively. The figure shows that

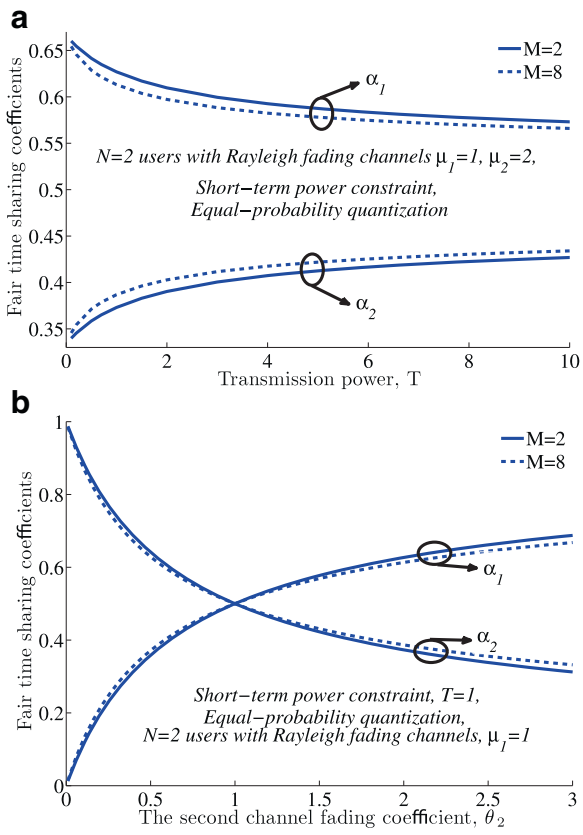

Figure 9 Optimal time-sharing coefficients versus (a) transmission power *T* and (b) the second user fading coefficient μ_2. $N = 2$ users, short-term power constraint, equal-probability fading quantization. The time-sharing parameters are determined such that the two users have the same throughput. M is the number of QRs considered for each user.

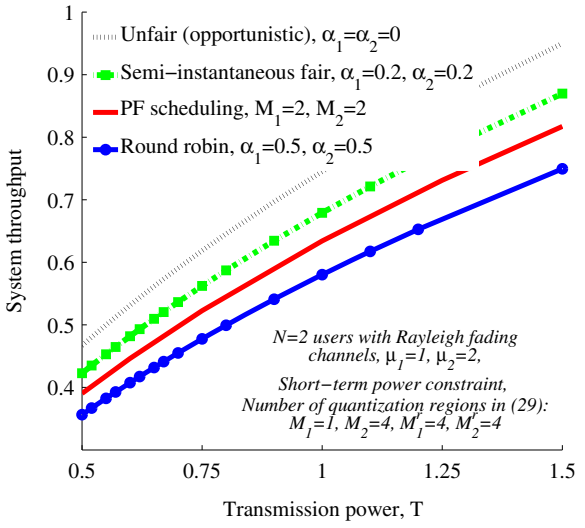

Figure 10 System throughput under different semi-instantaneous fairness constraints. The throughput is found according to (29) for $N = 2$ users with Rayleigh fading channels, $\mu_1 = 1, \mu_2 = 2, M_1 = 1, M_2 = 4, M_1^r = 4$, and $M_2^r = 4$. Short-term power constraint. The results are the same as the ones obtained by $M_1^r = 5, M_2^r = 5$ under extremely hard outage probability constraint (Theorem 4).

- Considering a long-term fairness constraint, no data transmission is allowed at low transmission powers, as the worse user (user 1) individual throughput constraints, i.e., (23.II), can not be satisfied.
- Further, even if the weak users' long-term fairness constraints are satisfied, the sum throughput is less than the one with no fairness constraint (please compare the sum throughput of the unfair scheduling curves and the ones with long-term fairness constraint). This is due to the fact that part of the throughput is sacrificed to satisfy the users' fairness requirements.
- Assuming a long-term fairness constraint, the optimal scheduling protocol is completely different from the one expected in the common schedulers. For instance, with the parameter setting of Figure 8, the worse user (user 1) individual throughput constraint is satisfied with less transmission power in the $M_1 = 1$, $M_2 = 4$ case, when compared with the $M_1 = M_2 = 2$ case. This is because, although less partial CSI is allocated to user 1 in the first case, it is more often scheduled, i.e., its scheduling probability in (23.II) increases.
- To increase the sum throughput, more feedback resources should be allocated to users experiencing better channels, even if there is a fairness constraint (please compare the cases $M_1 = 1$, $M_2 = 4$, and $M_1 = M_2 = 2$).
- Finally, the system performance is highly affected by the bit allocation procedure. This is a motivation for investigating optimal bit allocation in the throughput-fairness problem in the future.

5.6 The system throughput with time-sharing and a semi-instantaneous fairness constraint

The system performance under semi-instantaneous fairness constraint is investigated in Figures 9 and 10. Considering two users with different fading coefficients, Figure 9 shows the optimal time-sharing parameters such that the two users have the same throughput. That is, the time-sharing coefficients are determined such that $\alpha_1\bar{R}(1) = \alpha_2\bar{R}(2)$. Here, it is focused on short-term power constraint and equal-probability quantizers. Further, Figure 10 studies the effect of the different semi-instantaneous fairness constraints on the system throughput. Here, the results are obtained according to (29) for $N = 2$ users with fading coefficients $\mu_1 = 1$, $\mu_2 = 2$, $M_1 = 1$, $M_2 = 4$, $M_1^r = 4$, and $M_2^r = 4$ number of QRs in the opportunistic and round robin-based time slots. Note that the results

are the same as the ones obtained by $M_1 = 2$, $M_2 = 5$, $M_1^{\mathrm{r}} = 5$, and $M_2^{\mathrm{r}} = 5$ under extremely hard outage probability constraint (Theorem 4, Remark 5). Finally, Figures 8 and 10 compare the results with the achievable throughput in the standard PF scheduling scheme, where the users are scheduled based on their normalized SNR. Here, the following results are deduced from the figures:

- Assuming a semi-instantaneous fairness constraint, the sum throughput decreases considerably as the users' fairness constraints get harder (Figure 10).
- Setting the time-sharing coefficients to have fair scheduling between the receivers, the coefficients get closer together when the transmission power or the number of QRs increases (Figure 9).
- Finally, depending on the fairness constraints and the transmission power, higher throughput can be achieved by either PF scheduling or the proposed methods (Figures 8 and 10). However, our schemes are more flexible in tackling the throughput-fairness tradeoff, as the users are weighted and not necessarily equalized.

6 Conclusion

This paper addressed the fairness, power allocation and CSI quantization problems in the multiuser networks using multiple feedback bits per user. The analytical and simulation results show that the system outage probability vanishes when the number of users increases. At high SNRs, the optimal quantization boundaries get closer to zero. Optimal CSI quantization highly affects the system throughput specifically when the number of users increases or the transmission power decreases. The users' hard outage probability constraints can be satisfied at the cost of one QR. The throughput-fairness tradeoff can be properly addressed by combination of different scheduling procedures. However, the long-term or semi-instantaneous fairness constraints reduce the system throughput substantially. Using optimal power allocation, the first QR of each user is the only region for which no power may be allocated. Moreover, the first QR is the only region where the outage may occur. Finally, the feedback bit allocation in the throughput-fairness problem is a challenging issue for future works.

Endnotes

[a] Due to high number of papers dealing with partial CSI and multiuser diversity, it is not possible to mention all related works here. We apologize to the authors whose papers we have not included in our list and refer the readers to references in [1-19,21-37] for deeper review of the related works.

[b] The term Z_n represents the Gaussian interference received from other transmitters/cells as well.

[c] Note that the rate $R(g_1, \ldots, g_N)$ includes both scheduling one of the good users and then transmission to that user. Therefore, it is a function of the channel gains to all users.

[d] With no fairness constraints and equal power allocation, the optimal user selection metric maximizing the throughput would be the expected achievable rate of the QRs.

[e] Note that although scheduling is affected by the fairness constraint, the property $\varphi_{n,m} \leq \varphi_{n,m+1}$ is still valid as it only deals with two successive QRs in an individual user.

Appendices

Appendix 1

Proof of Theorem 1

The inequality (12.I) is obtained by (8), removing the constant 1 from the logarithmic terms and using the *log sum inequality* $\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq (\sum_{i=1}^{n} a_i) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$ ([51], eq. 2.165). For the log sum inequality, we replace a_i and b_i by $\varphi_{n,m}\beta_{n,m}$ and $\frac{\varphi_{n,m}\beta_{n,m}}{\hat{g}_{n,m}T_{n,m}}$, respectively. Also, (12.II) is based on the concavity of the function $f(x) = \log(1+x)$, Jensen's inequality, i.e., $f(E(x)) \geq E(f(x))$ for concave functions [51], and taking expectation with respect to probability terms $\{\frac{\varphi_{n,m}\beta_{n,m}}{\sum_{n=1}^{N}\sum_{m=0}^{M_n}\varphi_{n,m}\beta_{n,m}}\}$.

Appendix 2

Proof of Theorem 3

Considering Figure 1a, the arguments are easily trackable. Suppose that for a given scheduling policy, the quantization parameters $\tilde{g}_{n,m}$ and $\hat{g}_{n,m}, n = 1, \ldots, N$, $m = 1, \ldots, M_n$, have been optimized, in terms of (8), and we have $\tilde{g}_{n,m} \neq \hat{g}_{n,m+1}$. Provided that nth user being in the $(m+1)$th region has been selected, the codeword is sent at rate $R_{n,m+1} = \log(1 + \hat{g}_{n,m+1}T)$. If the instantaneous gain realization is within the region $\tilde{g}_{n,m} \leq g_n < \hat{g}_{n,m+1}$ (the outage region in Figure 1a), the data is lost, and the transmitted information has no contribution on the system throughput. However, by considering this interval as a part of the mth QR, while keeping the scheduling policy the same as before, we increase the system throughput by $\varphi_{n,m}(F_{G_n}(\hat{g}_{n,m+1}) - F_{G_n}(\tilde{g}_{n,m}))\log(1+\hat{g}_{n,m}T)$. This is because of the fact that as $\hat{g}_{n,m} < \tilde{g}_{n,m} \leq g_n$, the new codeword transmitted at rate $\log(1 + \hat{g}_{n,m}T)$ is definitely decoded by the receiver. The rate increment is in contrast to our first optimality assumption. Therefore, it is concluded that in the optimal case, we have $\tilde{g}_{n,m} = \hat{g}_{n,m+1}, \forall n, m$. Finally, note that in [43], the same argument has been proved for single-user networks. However, the gradient-based arguments of [43] are not implementable here, because there is no closed-form expression for the scheduling procedure.

Appendix 3

Proof of Theorem 4

Setting $\xi_n \to 0$, the outage probability constraint $\hat{g}^{\mathrm{r}}_{n,1} \leq F^{-1}_{G_n}(\xi_n)$ implies $\hat{g}^{\mathrm{r}}_{n,1} \to 0$. Therefore, with a short-term power constraint, the RR-based contribution of the nth user on the system throughput in (29) is obtained by

$$\lim_{\xi_n \to 0} \bar{\theta}(n|M^{\mathrm{r}}_n, \xi_n) = \max_{\tilde{g}^{\mathrm{r}}_{n,m}, \hat{g}^{\mathrm{r}}_{n,m}} \sum_{m=2}^{M^{\mathrm{r}}_n} \left(F_{G_n}(\tilde{g}^{\mathrm{r}}_{n,m}) - F_{G_n}(\hat{g}^{\mathrm{r}}_{n,m})\right) \times \log(1+\hat{g}^{\mathrm{r}}_{n,m}T), \hat{g}^{\mathrm{r}}_{n,m} = \tilde{g}^{\mathrm{r}}_{n,m+1} \forall n,m \geq 1. \tag{30}$$

However, defining $\hat{q}^{\mathrm{r}}_{n,1} = \tilde{g}^{\mathrm{r}}_{n,1}$ and $\tilde{q}^{\mathrm{r}}_{n,m} = \tilde{g}^{\mathrm{r}}_{n,m+1}, m \geq 1$, (30) can be rewritten as

$$\max_{\tilde{q}^{\mathrm{r}}_{n,m}, \hat{q}^{\mathrm{r}}_{n,m}} \sum_{m=1}^{M^{\mathrm{r}}_n - 1} \left(F_{G_n}(\tilde{q}^{\mathrm{r}}_{n,m}) - F_{G_n}(\hat{q}^{\mathrm{r}}_{n,m})\right) \log(1 + \hat{q}^{\mathrm{r}}_{n,m}T), \quad \hat{q}^{\mathrm{r}}_{n,m} = \tilde{q}^{\mathrm{r}}_{n,m+1} \, \forall n, m \geq 1 \tag{31}$$

which is the RR-based throughput with no outage probability constraint and $M^{\mathrm{r}}_n - 1$ QRs, i.e., $\bar{\theta}(n|M^{\mathrm{r}}_n - 1, 1)$. Therefore, as under short-term power constraint, the parameters $\tilde{g}^{\mathrm{r}}_{n,m}$ and $\hat{g}^{\mathrm{r}}_{n,m}$ do not affect the other terms of the optimization problem (29), we have $\lim_{\xi_n \to 0} \bar{R}^{\mathrm{semi}}_{\max}(1, \ldots, N|M^{\mathrm{r}}_n, \xi_n) = \bar{R}^{\mathrm{semi}}_{\max}(1, \ldots, N|M^{\mathrm{r}}_n - 1, 1)$, as stated in the theorem.

Competing interests

The authors declare that they have no competing interests.

Received: 8 January 2013 Accepted: 29 September 2013
Published: 24 October 2013

References

1. T Eriksson, T Ottosson, Compression of feedback in adaptive OFDM-based systems using scheduling. IEEE Commun. Lett. **11**(11), 859–861 (2007)
2. T Eriksson, T Ottosson, Compression of feedback for adaptive transmission and scheduling. Proc. IEEE. **95**(12), 2314–2321 (2007)
3. DJ Love, RW Heath, VKN Lau, D Gesbert, BD Rao, M Andrews, An overview of limited feedback in wireless communication systems. IEEE J. Sel. Areas Commun. **26**(8), 1341–1365 (2008)
4. A Haghighat, Z Lin, G Zhang, Haar compression for efficient CQI feedback signaling in 3GPP LTE systems. Paper presented at the IEEE wireless communications and networking conference, Las Vegas, NV, 31 March–3 April 2008
5. B Makki, T Eriksson, Efficient channel quality feedback signaling using transform coding and bit allocation. Paper presented at the IEEE 71st vehicular technology conference, Taipei, 16–19 May 2010
6. R1-070368, *System level comparison of best-M and DCT-based CQI compression schemes*. (Huawei, RAN1 meeting 47bis, Sorrento, Italy, 2007). available at http://www.3gpp.org
7. R1-063086, *Overhead reduction of best-M based CQI reporting*. (Huawei, RAN1 meeting 47, Riga, Latvia, 2006). available at http://www.3gpp.org
8. S Sanayei, A Nosratinia, Exploiting multiuser diversity with only 1-bit feedback. WCNC. **2**, 978–983 (2005)
9. VKN Lau, Proportional fair space-time scheduling for wireless communications. IEEE Trans. Commun. **53**(80), 1353–1360 (2005)
10. F Boccardi, F Tosato, G Caire, Precoding schemes for the MIMO-GBC. Paper presented at the International Zurich Seminar on Communications, Zurich, 22–24 February 2006
11. A Wiesel, YC Eldar, S Shamai, Zero-forcing precoding and generalized inverses. IEEE Trans. Sig. Proc. **56**(9), 4409–4418 (2008)
12. H Nam, M-S Alouini, Multiuser switched diversity scheduling systems with per-user threshold. IEEE Trans. Commun. **58**(5), 1321–1326 (2010)
13. K-K Wong, J Chen, Near-optimal power allocation and multiuser scheduling with outage capacity constraints exploiting only channel statistics. IEEE Trans. Wireless Commun. **7**(3), 812–818 (2008)
14. YS Al-Harthi, AH Tewfik, M-S Alouini, Multiuser diversity with quantized feedback. IEEE Trans. Wireless Commun. **6**(1), 330–337 (2007)
15. H Zhou, P Fan, D Guo, Joint channel probing and proportional fair scheduling in wireless networks. IEEE Trans. Wireless Commun. **10**(10), 3496–3505 (2011)
16. S Schwarz, C Mehlfuhrer, M Rupp, Throughput maximizing multiuser scheduling with adjustable fairness. Paper presented in IEEE international conference on communications, Kyoto, 5–9 June 2011
17. CW Park, H-J Lee, J-T Lim, Capacity and fairness trade-off in an outage situation over multiuser diversity systems. IEEE Commun. Lett. **15**(2), 184–186 (2011)
18. KK Wong, J Chen, Time-division multiuser MIMO with statistical feedback. EURASIP J. Adv. Sig. Proc. **2008**, 632134 (2008)
19. F Floren, O Edfors, B-A Molin, The effect of feedback quantization on the throughput of a multiuser diversity scheme. GLOBECOM. **1**, 497–501 (2003)
20. WiMAX Forum, *Mobile WiMAX - Part I: A Technical Overview and Performance Evaluation*. Tech. Rep. (WiMAX Forum, Clackamas, OR , 2006), pp. 1–53
21. D Gesbert, MS Alouini, Selective multi-user diversity. Paper presented in the proceedings of the 3rd IEEE international symposium on signal processing and information technology, Tucson, USA, 14–17 December 2003
22. CS Hwang, K Seong, J Cioffi, Throughput maximization by utilizing multi-user diversity in slow-fading random access channels. IEEE Trans. Wireless Commun. **7**(7), 2526–2535 (2008)
23. W Zhang, KB Letaief, MIMO broadcast scheduling with limited feedback. IEEE J. Sel. Areas Commun. **25**(7), 1457–1467 (2007)
24. SY Jeon, DH Cho, An enhanced channel-quality indication (CQI) reporting scheme for HSDPA systems. IEEE Commun. Lett. **9**(5), 432–434 (2005)
25. A Kuhne, A Klein, Throughput analysis of multi-user OFDMA-systems using imperfect CQI feedback and diversity techniques. IEEE J. Sel. Areas Commun. **26**(8), 1440–1450 (2008)
26. YS Al-Harthi, Opportunistic multiuser scheduling with reduced feedback load. Eur. Trans. Telecomms. **21**, 299–311 (2010)
27. H Nam, YC Ko, M-S Alouini, Performance analysis of joint switched diversity and adaptive modulation. IEEE Trans. Wireless Commun. **7**(10), 3780–3790 (2008)
28. L Li, M Pesavento, AB Gershman, Downlink opportunistic scheduling with low-rate channel state feedback: Error rate analysis and optimization of the feedback parameters. IEEE Trans. Commun. **58**(10), 2871–2880 (2010)
29. APT Lau, FR Kschischang, Feedback quantization strategies for multiuser diversity systems. IEEE Trans. Info. Theory. **53**(4), 1386–1400 (2007)
30. K Huang, JG Andrews, RW Heath, Performance of orthogonal beamforming for SDMA with limited feedback. IEEE Trans. Veh. Tech. **58**(1), 152–164 (2009)
31. J Zhang, M Kountouris, JG Andrews, RW Heath, Achievable throughput of multi-mode multiuser MIMO with imperfect CSI constraints. Paper presented in IEEE international symposium on information theory, Seoul, 28 June–3 July 2009
32. B Khoshnevis, W Yu, Bit allocation laws for multiantenna channel feedback quantization: Multiuser case. IEEE Trans. Sig. Proc. **60**(1), 367–382 (2012)
33. AG Marques, GB Giannakis, J Ramos, Optimizing orthogonal multiple access based on quantized channel state information. IEEE Trans. Sig. Proc. **59**(10), 5023–5038 (2011)
34. T Kim, JT Lim, Capacity and fairness tradeoff in multiuser scheduling system with reduced feedback. IEEE Commun. Lett. **13**(11), 841–843 (2009)
35. JW So, JM Cioffi, Feedback reduction scheme for downlink multiuser diversity. IEEE Trans. Wireless Commun. **8**(2), 668–672 (2009)
36. JW So, JM Cioffi, Capacity and fairness in multiuser diversity systems with opportunistic feedback. IEEE Commun. Lett. **12**(9), 648–650 (2008)

37. Y Soydan, C Candan, A feedback quantization scheme leveraging fairness and throughput for heterogeneous multi-user diversity systems. IEEE Trans. Veh. Tech. **59**(5), 2610–2614 (2010)
38. J Chen, KK Wong, An energy-saving QoS-based resource allocation for multiuser TDMA systems with causal CSI. Paper presented at the IEEE global telecommunications conference, New Orleans, 30 November–4 December 4 2008
39. J Chen, KK Wong, Multiuser MIMO-TDMA with statistical feedback. Paper presented at the 6th international conference on information, communications and signal processing, Singapore, 10–13 December 2007
40. R Knopp, PA Humblet, Information capacity and power control in single-cell multiuser communications. ICC. **1**, 331–335 (1995)
41. DN Tse, Optimal power allocation over parallel gaussian broadcast channels. Paper presented at the IEEE international symposium on information theory, Ulm, 29 June–4 July 1997
42. G Caire, G Taricco, E Biglieri, Optimum power control over fading channels. IEEE Trans. Info. Theory. **45**(5), 1468–1489 (1999)
43. TT Kim, M Skoglund, On the expected rate of slowly fading channels with quantized side information. IEEE Trans. Commun. **55**(4), 820–829 (2007)
44. B Makki, T Eriksson, On the average rate of quasi-static fading channels with ARQ and CSI feedback. IEEE Commun. Lett. **14**(9), 806–808 (2010)
45. B Makki, T Eriksson, On the average rate of HARQ-based quasi-static spectrum sharing networks. IEEE Trans. Wireless Commun. **11**(1), 65–77 (2012)
46. B Makki, T Eriksson, Data transmission in the presence of noisy channel state feedback and outage probability constraint. Paper presented at the international symposium on information theory and its applications. Taichung, 17–20 October 2010
47. G Caire, D Tuninetti, The throughput of hybrid-ARQ protocols for the Gaussian collision channel. IEEE Trans. Info. Theory. **47**(5), 1971–1988 (2001)
48. S Tatikonda, S Mitter, The capacity of channels with feedback. IEEE Trans. Info. Theory. **55**(1), 323–349 (2009)
49. A Ephremides, B Hajek, Information theory and communication networks: An unconsummated union. IEEE Trans. Info. Theory. **44**(6), 2416–2434 (1998)
50. B Makki, M Noori Hosseini, SA Seyyedsalehi, N Sadati, Unaligned training for voice conversion based on a local nonlinear principal component analysis approach. Neural Comp. App. **19**(3), 437–444 (2009)
51. TM Cover, JA Thomas, *Elements of Information Theory* (Wiley Interscience, New York, 2006), pp. 776

doi:10.1186/1687-1499-2013-249
Cite this article as: Makki and Eriksson: **Fairness, power allocation, and CSI quantization in block fading multiuser systems.** *EURASIP Journal on Wireless Communications and Networking* 2013 **2013**:249.