Beam Selection Strategies for Orthogonal Random Beamforming in Sparse Networks

Jose Lopez Vicario, *Member, IEEE,* Roberto Bosisio, Carles Anton-Haro, *Senior Member, IEEE,* and Umberto Spagnolini, *Senior Member, IEEE*

***Abstract*—Orthogonal random beamforming (ORB) constitutes a mean to exploit spatial multiplexing and multi-user diversity (MUD) gains in multi-antenna broadcast channels. To do so, as many random beamformers as transmit antennas (M) are generated and on each beam the user experiencing the most favorable channel conditions is scheduled. Whereas for a large number of users the sum-rate of ORB exhibits an identical growth rate as that of dirty paper coding, performance in sparse networks (or in networks with an uneven spatial distribution of users) is known to be severely impaired. To circumvent that, in this paper we modify the scheduling process in ORB in order to select a *subset* out of the M available beams. We propose several beam selection algorithms and assess their performance in terms of sum-rate and aggregated throughput (i.e., rate achieved with practical modulation and coding schemes), along with an analysis of their computational complexity. Since ORB schemes require partial channel state information (CSI) to be fed back to the transmitter, we finally investigate the impact of CSI quantization on system performance. More specifically, we prove that most of the MUD can be still exploited with very few quantization bits and we derive a beam selection approach trading-off system performance vs. feedback channel requirements.**

***Index Terms*—Orthogonal Random Beamforming (ORB), beam selection, sparse networks, opportunistic scheduling, Multi-user Diversity (MUD), broadcast channel, feedback quantization.**

I. Introduction

In recent years, applications that involve the transmission over broadcast channels in multi-user wireless systems have increased the interests. In the context of single-input single-output (SISO) systems, the strategy that maximizes the sum-rate consists in selecting for each time slot the user which experiences the most favorable channel conditions [1], [2]. Dirty paper coding (DPC) constitutes the capacity-achieving strategy for the Gaussian multiple-input multiple-output (MIMO) broadcast channel [3] and, consequently, it maximizes the sum-rate too [4]. In DPC, multiple users can be simultaneously served thanks to the spatial multiplexing capabilities of the MIMO links. However, DPC is not practical due to the need for successive encodings and decodings. Transmit zero-forcing (ZF) is an affordable suboptimal alternative as it exhibits the same sum-rate growth as DPC for an asymptotically high number of users when efficient user selection is adopted [5], [6]. Unfortunately, both ZF and DPC schemes require perfect channel state information (CSI) at the transmitter, which is seldom available. Instead, the so-called orthogonal random beamforming (ORB) schemes [7] only demand partial CSI at the base station (BS). In ORB, the transmitted data is pre-coded with a set of randomly-generated beamformers for which each user has to report its instantaneous signal-to-interference-and-noise ratio (SINR). This leads to a substantial decrease of the amount of data to be conveyed over the feedback channel but, still, the sum-rate exhibits an identical growth rate to that of the DPC and ZF schemes for an asymptotic number of users [7]. However, the performance of ORB for a realistic number of users is far from being satisfactory.

Contributions: In this paper, we investigate scheduling strategies for ORB suitable for sparse networks. When the power is evenly allocated to transmit antennas we show that having as many active beams as the number of transmit antennas is not always optimal. Instead, it is more appropriate to select a subset of active beams according to propagation conditions (i.e., for noise- or interference-limited scenarios) and the objective function (sum-rate or aggregated throughput). We investigate some beam selection algorithms featuring different complexity levels and performance gaps with respect to a brute-force approach. All those schemes merely require partial CSI at the transmitter (either the measured SINRs or, alternatively, the channel gains associated to each beam). In addition, we derive a closed-form expression of the aggregated throughput in the presence of adaptive coding and modulation schemes and, thus, we extend the work in [8] to the multiple antenna case. Such an expression is obtained both for the cases of analog or quantized partial CSI. As observed in the single-antena case [8], [9], we find out that most of the MUD gain can be preserved even when the measured SINRs are roughly quantized with a very few bits. As for the quantization process, we empirically prove that a non-uniform law based on the post-scheduling SINR distribution is an efficient choice to exploit MUD.

Relation to prior work: The optimization of ORB for sparse networks is an open issue addressed by few authors.

Manuscript received October 5, 2006; revised January 18, 2007, October 16, 2007, and March 27, 2008; accepted July 15, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was X. Wang. This work was partially supported by the European Commission under projects NEWCOM++ (216715) and COOPCOM (033533); and the Catalan Government (DURSI) 2005FI 00003, SGR2005-00690.

J. L. Vicario was with the Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), 08860 Castelldefells, Spain. He is now with the Universitat Autonoma de Barcelona (UAB), 08193 Bellaterra, Spain (e-mail: vicario@ieee.org).

C. Antón-Haro is with the Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), 08860 Castelldefells, Spain (e-mail: carles.anton@cttc.es).

R. Bosisio and U. Spagnolini are with the Politecnico de Milano, I-20133 Milano, Italy (e-mail: bosisio.roberto@gmail, spagnoli@elet.polimi.it).

Digital Object Identifier 10.1109/TWC.2008.060794.

Kountouris and Gesbert proposed to exploit the temporal correlation of the channel response in order to identify the best subset of beamformers [10]. In [11], the same authors proposed a method to make performance less sensitive to channel properties. By means of a low-rate feedback channel the most promising user subset is selected first. Then, more sophisticated techniques are applied to such user subset according to the nature of the CSI at the transmitter (full or partial). Among these techniques, it was proposed an algorithm which motivated the idea of embedding a beam selection mechanism into ORB. According to the SINR of the different users, the transmission scheme switched from SDMA (all the available beams active) to TDMA (only the best beam active). Kobayashi *et al.* also showed that activating all the available beams is not the optimal choice when the feedback channel is delayed [12].

Independently from our work (part of it initially presented in [13], [14]), the pure concept of beam selection has been recently investigated and further developed in [15] and [16]. Our contribution here is more extensive as we compare different beam selection methods and consider the inherent trade-offs in terms of system performance vs. the amount of information required in the feedback channel. Furthermore, we show that most of the MUD provided by ORB can be exploited with a few bits in the feedback channel and we derive a beam selection approach trading-off performance vs. feedback bits requirements. Our study is then more practical than that carried out in [17], where it is shown that most of MUD gain in ORB can be extracted with one feedback bit when the number of users is high enough (say $K \geq 1000$ users).

Organization: In Section II, the corresponding signal model is presented. Then, in Section III, we derive closed-form expressions for both the probability and cumulative density functions (pdf and CDF) of the post-scheduling SINRs. In Sections IV and V, we assess ORB performance both in terms of sum-rate and aggregated throughput, respectively. Next, in Section VI, we propose several beam selection algorithms accompanied by a system performance vs. complexity analysis. In Section VII we evaluate the impact of feedback quantization on the aggregated throughput. Finally, in Section VIII, we conclude by summarizing the main results obtained in this paper.

II. Signal Model

Consider the downlink of a wireless system with one BS equipped with M antennas and K single-antenna mobile stations (MS). In order to serve multiple users in the same time-slot, a linear precoding matrix is applied at the base station. According to the orthogonal random beamforming (ORB) strategy [7], in each time-slot we generate a random matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M]$, where $\mathbf{w}_i \in \mathbb{C}^{M\times 1}$, $i = 1, ..., M$, are random orthonormal vectors drawn from an isotropic distribution [18]. Then, each of those vectors are used for transmitting different streams to the users with the highest SINRs. Unlike the approach in [7], here we do not necessarily transmit with all the beams vectors $\mathbf{w}_i$, but rather, the transmission is made with a subset of active beams $\mathcal{B} \subset \mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$ (details on the beam selection procedure are given in Section VI). The received signal at the k-th MS can be expressed as:

$$r_k = \mathbf{h}_k^T \mathbf{W}_{\mathcal{B}} \mathbf{s}_{\mathcal{B}} + n_k \tag{1}$$

where $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ is the channel vector gain between the BS and the k-th MS modelled as $\mathbf{h}_k \sim \mathcal{CN}(0, \mathbf{I}_M)$ (independent Rayleigh fading), $\mathbf{W}_{\mathcal{B}} \in \mathbb{C}^{M\times B}$ is the precoding matrix with the columns of $\mathbf{W}$ corresponding to the subset of active beams $\mathcal{B}$, $\mathbf{s}_{\mathcal{B}} \in \mathbb{C}^{B\times 1}$ is the symbol vector broadcasted from the BS with $B = \text{card}(\mathcal{B}) \leq M$ active beams, and where $n_k \sim \mathcal{CN}(0, \sigma^2)$ is AWGN. The active users in the system are assumed to undergo independent Rayleigh fading processes. Further, we consider block-fading where the channel response remains constant during one time-slot and changes to a new independent realization in the subsequent one. Concerning channel state information, we assume perfect CSI knowledge of $\mathbf{h}_k$ for *each* user at the receive side, and the availability of a low-rate error-free feedback channel to convey partial CSI to the transmitter. The total transmit power is constant and evenly distributed among the active beams so that $\mathbb{E}\{\mathbf{s}_{\mathcal{B}}^H \mathbf{s}_{\mathcal{B}}\} = P_t$ and, thus, each active beam is allocated a transmit power equal to P_t/B. The average signal-to-noise ratio (SNR) of the system is defined as $\rho = \frac{P_t}{\sigma^2}$.

III. Post-scheduling SINR Statistics

Before deriving the beam selection algorithms, we analyze the system in a context where the number of active beams B is kept constant but still $B \leq M$. According to the signal model of the previous section, the received signal for user k when using beamformer $\mathbf{w}_i$ can be re-written from (1) as:

$$r_k = \mathbf{h}_k^T \mathbf{w}_i s_i + \sum_{j\in\mathcal{B}, j\neq i} \mathbf{h}_k^T \mathbf{w}_j s_j + n_k \tag{2}$$

where s_j stands for the symbol transmitted with beam j. Notice that the last two terms in the above expression are associated with the interference-plus-noise contribution and, hence, the corresponding SINR for k-th user and i-th beam amounts to

$$\text{SINR}_{k,i} = \frac{|\mathbf{h}_k^T \mathbf{w}_i|^2}{B/\rho + \sum_{j\in\mathcal{B}, j\neq i} |\mathbf{h}_k^T \mathbf{w}_j|^2} = \frac{z}{B/\rho + y} \,. \tag{3}$$

Since we assume that all users experience i.i.d Rayleigh fading and the active beams are orthonormal to each other, the two random variables z and y are independent chi-square distributed: $z \sim \chi^2_2$ and $y \sim \chi^2_{2B-2}$. Bearing this in mind, the CDF and pdf of the SINR can be expressed as [7]:

$$F_{\text{SINR}}(\gamma) = 1 - \frac{e^{-\gamma B/\rho}}{(1+\gamma)^{B-1}} \tag{4}$$

$$f_{\text{SINR}}(\gamma) = \frac{e^{-\gamma B/\rho}}{(1+\gamma)^{B}} \left(\frac{B}{\rho}(1+\gamma) + B - 1\right) \tag{5}$$

Notice that in a i.i.d Rayleigh fading scenario the SINR statistics only depend on the *number of active beams* B.

The scheduling process is organized in a slot-by-slot basis following a *max*-SINR rule. That is, each user reports to the BS the SINR associated to its best beam and the beam index. With this information, the BS selects for each beam the user with the highest SINR. By assuming that a different user is

chosen for each beam, the active user k_i^* selected for beam i is then that satisfying[1] $k_i^* = \arg\max_{k=1,\ldots,K} \{\mathsf{SINR}_{k,i}\}$. As shown in Section VI, however, CSI and user selection is given in a different form when beam selection procedures are in use.

Since all the users experience i.i.d Rayleigh fading, the CDF of the SINR experienced by the scheduled user (the *post-scheduling* SINR) $F_{\mathsf{SINR}^*}(\gamma)$ can be readily expressed as [7]:

$$F_{\mathsf{SINR}^*}(\gamma) = (F_{\mathsf{SINR}}(\gamma))^K = \left(1 - \frac{e^{-\frac{\gamma B}{\rho}}}{(1+\gamma)^{B-1}}\right)^K$$

Finally, by simply differentiating the above expression the corresponding pdf of post-scheduling SINR follows:

$$f_{\mathsf{SINR}^*}(\gamma) = K\frac{e^{-\frac{\gamma B}{\rho}}}{(1+\gamma)^B}\left(\frac{B}{\rho}(1+\gamma)+B-1\right) \times \left(1 - \frac{e^{-\frac{\gamma B}{\rho}}}{(1+\gamma)^{B-1}}\right)^{K-1} \quad (6)$$

IV. Sum-rate Analysis

Here, we evaluate the ORB performance as a function of the number of active beams B. The purpose is to analyze whether activating all the available beams is an appropriate strategy in scenarios with a low number of users, since in those scenarios the probability that the generated random beamforming vectors match users' channel characteristics is considerably lower. We begin our study by showing the behavior of the system in terms of the sum-rate. According to the proposed scheduling policy, the achievable sum-rate when B beams are active is given by[2]:

$$R(\mathcal{B}) \simeq \mathbb{E}_\gamma\left[\sum_{i\in\mathcal{B}} \log_2\left(1 + \max_{1\le k\le K} \mathsf{SINR}_{k,i}(\mathcal{B})\right)\right] = B\int_{\gamma=0}^{\infty} \log_2(1+\gamma)\, f_{\mathsf{SINR}^*}(\gamma)d\gamma \quad (7)$$

Sharif and Hassibi [7] derived a closed-form expression for the asymptotic case ($K \to \infty$) which exhibits the same sum-rate growth as DPC. For a practical scenario with a finite number of users, though, resorting to numerical integration is needed. Still, this expression is tractable when the average SNR of the system is arbitrarily high ($\rho\to\infty$) and the system is so called interference dominated ($B>1$). In this case, the pdf of the post-scheduling SINR given by Eq. (6) can be re-written as follows:

$$f_{high,\mathsf{SINR}^*}(\gamma) = K\frac{B-1}{(1+\gamma)^B}\left(1 - \frac{1}{(1+\gamma)^{B-1}}\right)^{K-1} \quad (8)$$

As a consequence, a closed-form expression can be readily derived for the sum-rate [19]:

$$R_{high} \simeq \frac{B}{B-1}\log_2(e)\sum_{k=1}^{K}\frac{1}{k}\ ; \qquad B>1 \quad (9)$$

where the term $\sum_{k=1}^{K}\frac{1}{k}$ accounts for the multi-user gain. Clearly, the sum-rate decreases with the number of active beams due to the $\frac{B}{B-1}$ term, which suggests that a reduced number of beams should be used in interference-limited scenarios as also stated in [7], [15].

A different scenario results in the low-SNR regime. By neglecting the interference term in Eq. (3) (i.e., $B/\rho + y \to B/\rho$ for $\rho \to 0$), the post-scheduling pdf reads

$$f_{low,\mathsf{SINR}^*}(\gamma) = \frac{K}{\rho}e^{-\frac{\gamma B}{\rho}}\left(1 - e^{-\frac{\gamma B}{\rho}}\right)^{K-1}.$$

Notice that this expression is identical to that of a multi-user system with a SISO configuration and average $\mathsf{SNR} = \rho/B$ [20]. This follows from the fact that we assume i.i.d Rayleigh channel fading and we generate the orthonormal vectors $\mathbf{w}_i$ according to an isotropic distribution. Then, one can easily show that the sum-rate can be expressed as [21]:

$$R_{low} \simeq BK\log_2 e \times \sum_{k=0}^{K-1}\binom{K-1}{k}\frac{(-1)^{k+1}}{k+1}e^{B(k+1)/\rho}E_i\left(-B\frac{(k+1)}{\rho}\right) \quad (10)$$

where the exponential integral function is defined as $E_i(-x) \triangleq -\int_x^\infty \frac{e^{-t}}{t}dt$, for $x>0$ and can be written in series form as $E_i(-x) = e^{-x}\sum_{l=1}^{n}(-1)^l\frac{(l-1)!}{x^l} + R_n$, being R_n a remainder term [22]. By considering the above expression and bearing in mind that $\rho\to 0$, we can further simplify (10):

$$R_{low} \simeq BK\log_2 e\sum_{k=0}^{K-1}\binom{K-1}{k}\frac{(-1)^{k+1}}{k+1}\left(-\frac{\rho}{B(k+1)}\right) \simeq \rho\log_2(e)\sum_{k=1}^{K}\frac{1}{k} \quad (11)$$

where it is observed that the sum-rate still depends on the MUD gain but it is independent of B. It is worth noting that authors in [16] proved that the single-beam solution is the optimum choice in the case that $\rho\to 0$. In their proof, they considered a system adopting a beam selection algorithm. In this section, however, we have not already taken into consideration any beam selection strategy. Instead, we are investigating what is the most appropriate number of active beams when the original ORB scheme is used and, for that reason, a different result is obtained.

Table I and Fig. 1 illustrate the accuracy of the approximate sum-rate expressions for the low- and high-SNR regimes, respectively. On the one hand, Fig. 1 shows that the proposed high-SNR approximation becomes valid from ρ>25-30 dB. On the other hand, one can notice in Table I that when the SNR is low enough to neglect the interference term ($\rho = -5$ dB) both the simulated and approximated results given by Eq. (10) reflect the same trend for a growing number of beams. In the case that the SNR is considerably low ($\rho = -25$ dB),

[1]As proved in [7], the probability that one user achieves the highest SINR on more than one beam is negligible when the number of users is large compared with the number of active beams. In particular, the authors showed that the maximum number of beams ensuring a negligible probability for this event can be roughly expressed as $B\sim log_2 K$. In the case that a beam has not users who have reported their SINRs, a random user (out of the set of remaining users) is selected. However, this situation occurs with a low probability under the conditions of this paper ($B \le 4$ and $K \ge 10$) as the relation $B\sim log_2 K$ holds [7]. Besides, it has been empirically proven that the impact of this low probability event on system performance is negligible.

[2]The approximation comes from the fact that this expression is an upper bound of the sum-rate that ignores the probability that one user has the maximum SINR on more than one beam. As proved in [7], this upper bound is tight for those scenarios with a number of active beams lower or approximately equal to $B\sim log_2 K$ (as the scenario considered in this work).

TABLE I
SUM-RATE PERFORMANCE (BITS/S/HZ) IN LOW-SNR SCENARIOS (K=200 USERS).

	B=1	B=2	B=3	B=4
Simulation (ρ=-5 dB)	1.485	1.759	1.836	1.893
Eq. (10) (ρ=-5 dB)	1.485	1.848	2.071	2.179
Simulation (ρ=-25 dB)	0.0262	0.0267	0.0269	0.0267
Eq. (11) (ρ=-25 dB)	0.0268	0.0268	0.0268	0.0268

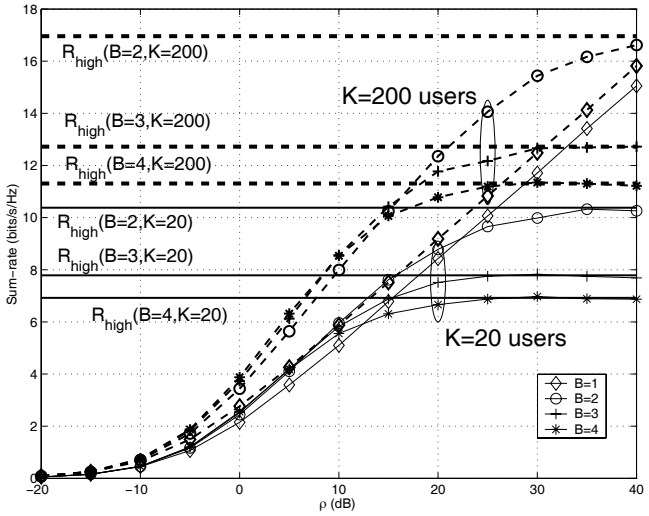

Fig. 1. Sum-rate vs. average SNR (ρ) for a different number of active beams. Horizontal lines correspond to the asymptotic results given by Eq. (9). Solid lines: K=20 users, dashed lines: K=200 users.

approximation in Eq. (11) becomes valid and, in this case, approximately the same sum-rate is achieved with different values of B. To summarize, in noise-limited scenarios a higher number of active beams turns out to be beneficial, in particular when K is high (see Fig. 1). Nonetheless, when the system becomes interference-limited or the SNR is extremely low, the use of multiple beams does not pay off. In the case of high SNR, activating a lower number of beams gives better results. In the sequel, we will consider ρ=0 dB as the low-SNR region. In our opinion, to take into consideration lower values of SNR is not useful for practical implementation. Notice that in this region using several beams is beneficial for sum-rate performance.

V. AGGREGATED THROUGHPUT ANALYSIS

Sum-rate measures obtained in the previous section simply provide a rough idea on how spectrally efficient the system can be when ideal Gaussian codebooks of large block length are available. However, in practical systems with a limited number of adaptive modulation and coding (AMC) modes and realistic coding methods, the achieved rate can be quite different. In that case, the aggregated link layer throughput provides a more realistic view.

A. Closed-form Expression

For a given modulation scheme, indexed by variable m, the aggregated throughput can be expressed as $\eta_m(\mathcal{B}) \simeq \mathbb{E}_\gamma\left\{\sum_{i\in\mathcal{B}} b_m\left(1-\mathrm{PER}_m(\max_{1\le k\le K}\mathsf{SINR}_{k,i})\right)\right\}$ where b_m is the number of bits per symbol and PER_m denotes the packet error probability. By considering L symbols packets, the aggregated throughput can be re-written as follows[3]:

$$\eta_m(\mathcal{B}) \simeq B\mathbb{E}_\gamma\left\{b_m(1-\mathrm{SER}_m(\max_{1\le k\le K}\mathsf{SINR}_{k,i}))^L\right\} = Bb_m\int_{\gamma=0}^{\infty}(1-\mathrm{SER}_m(\gamma))^L f_{\mathsf{SINR}^*}(\gamma)d\gamma \quad (12)$$

with SER_m standing for the symbol error rate associated to the modulation scheme m. As shown in [24], the SER for M-QAM modulation schemes can be approximated by:

$$\mathrm{SER}_m(\gamma) \simeq b_m 0.2\exp(-1.6\gamma/(2^{b_m}-1)) = \alpha_m\exp(-\beta_m\gamma) \quad (13)$$

where α_m and β_m are constellation-dependent parameters (being the approximation also quite accurate for BPSK by setting β_m=1). Note that, Eq. (13) is based on Gaussian approximation for the overall inter-user interference that holds when the number of interfering beams is high. Even if the proposed scheduler is aimed at finding the MSs which maximize the resulting SINR (or equivalently minimize inter-user interference), the interference term in Eq. (2) is expected to be small in comparison with the noise term, in particular when the number of users is high. Therefore, the Gaussian approximation proves to be quite accurate even for the B=2 case [21].

In the presence of adaptive modulation mechanisms, the aggregated throughput will depend on the modulation scheme selected for each beam (indexed by m_i) as follows:

$$\eta(\mathcal{B}) \simeq \mathbb{E}_\gamma\left\{\sum_{i\in\mathcal{B}} b_{m_i}(1-\mathrm{SER}_{m_i}(\max_{1\le k\le K}\mathsf{SINR}_{k,i}))^L\right\} \quad (14)$$

Concerning the adaptive modulation rule, we consider a cross-layer strategy that selects for each beam i the constellation size maximizing the instantaneous link layer throughput i.e., $m_i = \arg\max_{m\in\mathcal{M}} b_m(1-\mathrm{SER}_m(\gamma_{k^*,i}))^L$, where $\gamma_{k^*,i} = \max_{1\le k\le K}\mathsf{SINR}_{k,i}$ stands for the SINR corresponding to the scheduled user on beam i. From the above expression, the corresponding thresholds ($\gamma_{th,m}$) are obtained for a system with a number of modulation schemes given by the ordered set $\mathcal{M}$ = {BPSK,QPSK,16-QAM}. Consequently, the constellation size associated with the measured post-scheduling SINR on beam i is determined according to the rule $m_i = m \iff \gamma_{th,m} \le \gamma_{k^*,i} < \gamma_{th,m+1}$ with $\gamma_{th,1}$=0 and $\gamma_{th,\mathrm{card}(\mathcal{M})+1}$=$\infty$. As a result, the throughput for beam i can be computed as:

$$\mathbb{E}_\gamma\left\{b_{m_i}(1-\mathrm{SER}_{m_i}(\max_{1\le k\le K}\mathsf{SINR}_{k,i}))^L\right\} = \sum_{m=1}^{\mathrm{card}(\mathcal{M})} b_m\int_{\gamma=\gamma_{th,m}}^{\gamma_{th,m+1}}(1-\mathrm{SER}_m(\gamma))^L f_{\mathsf{SINR}^*}(\gamma)d\gamma$$

where one should bear in mind that the SINR of the different beams are identically distributed [7] and, for that reason, this

[3]For mathematical tractability, we restrict ourselves to uncoded transmissions but the analysis can be easily extended to transmissions with convolutional coding by resorting to the accurate exponential approximations of the PER derived in [23].

expression does not depend on index i. Finally, by introducing the previous result in (14) and by taking into account (13), one can re-writte the aggregated throughput in the presence of adaptive modulation as:

$$\eta(\mathcal{B}) \simeq B \sum_{m=1}^{\text{card}(\mathcal{M})} b_m \int_{\gamma=\gamma_{th,m}}^{\gamma_{th,m+1}} (1-\alpha_m e^{-\beta_m \gamma})^L f_{\mathsf{SINR}^*}(\gamma) d\gamma \tag{15}$$

After some algebraic manipulation (see Appendix), the aggregated throughput becomes:

$$\begin{aligned}\eta(\mathcal{B}) \simeq & \\ BK \sum_{m=1}^{\text{card}(\mathcal{M})} b_m \sum_{l=0}^{L} \binom{L}{l} (-\alpha_m)^l \sum_{k=0}^{K-1} \binom{K-1}{k} (-1)^k e^{\mu} \mu^c & \\ \times \left[\frac{B}{\rho\mu} \left(\Gamma\left(1-c, (1+\gamma_{th,m})\mu\right) - \Gamma\left(1-c, (1+\gamma_{th,m+1})\mu\right) \right) \right. & \\ \left. + (B-1) \left(\Gamma\left(-c, (1+\gamma_{th,m})\mu\right) - \Gamma\left(-c, (1+\gamma_{th,m+1})\mu\right) \right) \right] & \end{aligned} \tag{16}$$

where $\mu = \beta_m l + \frac{B}{\rho}(k+1)$, $c = (k+1)(B-1)$; and $\Gamma(\alpha, x) \triangleq \int_x^{\infty} e^{-t} t^{\alpha-1} dt$ stands for the complementary incomplete gamma function.

B. Asymptotic Analysis

In order to gain some insight, we analyze the behavior of the aggregated throughput in the asymptotic high SNR regime ($\rho \to \infty$). We focus on B=1 as this is the case that presents the main difference with respect to the sum-rate analysis. When B=1, c=0 and Eq. (16) can be re-written as:

$$\begin{aligned}\eta \simeq & \sum_{m=1}^{\text{card}(\mathcal{M})} b_m K \sum_{l=0}^{L} \binom{L}{l} (-\alpha_m)^l \\ & \times \sum_{k=0}^{K-1} \binom{K-1}{k} (-1)^k e^{\beta_m l + \frac{k+1}{\rho}} \frac{1}{\rho\beta_m l + k + 1} \\ & \times \left(e^{-(1+\gamma_{th,m})\left(\beta_m l + \frac{k+1}{\rho}\right)} - e^{-(1+\gamma_{th,m+1})\left(\beta_m l + \frac{k+1}{\rho}\right)} \right)\end{aligned}$$

where the equivalence $\Gamma(1, x) = e^{-x}$ has been used. When $\rho \to \infty$, all the terms in the summations with $l \neq 0$ vanish while for l=0 the summations also tend to zero except for the terms related to $m = \text{card}(\mathcal{M})$ since $\gamma_{th,\text{card}(\mathcal{M})+1} = \infty$. Then, η for $\rho \to \infty$ becomes

$$\begin{aligned}\eta_{high} \simeq & b_{\text{card}(\mathcal{M})} K \sum_{k=0}^{K-1} \binom{K-1}{k} \frac{(-1)^k}{k+1} \\ = & b_{\text{card}(\mathcal{M})} \sum_{t=1}^{K} \binom{K}{t} (-1)^{t-1} \\ = & b_{\text{card}(\mathcal{M})}\end{aligned}$$

Differently from the sum-rate case, a saturation effect is observed in the asymptotic SNR regime when B=1. This saturation effect (known as bit cap effect) is due to the fact that a finite number of practical modulation schemes are adopted [25]. The smaller the constellation size of the highest AMC level, the more severe the saturation effect is. For the case $B>1$ (i.e., $c \neq 0$) and $\rho \to \infty$, it can be easily shown that (as the sum-rate expression) Eq. (16) is monotonically decreasing in B (details are omitted here for brevity but can be found in [21]). Therefore, activating more than *two* beams is not beneficial in the high-SNR regime[4].

In conjunction with the previous asymptotic result, one can conclude that the most appropriate strategy could be using one or two active beams according with the size of the highest available AMC level. Using only one active beam may be the best option in the asymptotic regime when the constellation size of the highest AMC level is high and the number of users is low. In the opposite case, it is more appropriate to allocate the transmit power to two active beams as setting $B>2$ would be less efficient[5]. The numerical results shown in Fig. 2 confirm these conclusions. Since the single beam solution quickly saturates, increasing the number of beams pays off in the high-SNR region as long as the number of active beams is $B \leq 2$. Besides, in densely populated scenarios, the SINR associated with the active beams take higher values and, as a result, the gap between the curves corresponding B=1 and $B>1$ cases becomes wider. This is because interference generated by the beams tends to its average, whereas the power of the desired signal grows for increasing K. Notice that an asymptotic analysis cannot be carried out for low SNR as the SER approximation (13) is a loose approximation in this region [24]. Still, one can notice from the numerical results that using several beams is not an appropriate strategy for noise-limited scenarios as PER levels are very high to support several active beams, even with the lowest modulation scheme (BPSK). Spatial multiplexing capabilities can be exploited in the mid-SNR region of highly-populated cells (see Fig. 2), where the SNR is neither low enough to penalize PER behavior nor high enough to enter the interference-limited region.

In summary, due to granularity and saturation effects, different conclusions are drawn for aggregated throughput with respect to the sum-rate case. Nonetheless, we have proved in both cases that using all the active beams is not the best strategy in a scenario with a practical number of users (say $K < 100$). This motivates the need for developing beam selection schemes.

VI. Adaptive Beam Selection

In this section, we propose several beam selection strategies capable of identifying the best subset of beams (and users) according to scenario conditions. These algorithms are quite interesting for sparse networks as the value of K can be *virtually* increased. That is, the number of SINR *combinations* is larger and, then, system performance is improved as if K

[4]It is worth recalling that here we are not considering beam selection strategies.

[5]Analytically, it is not straightforward to obtain the values of K and modulation schemes for which it is better using one or two beams. Then, numerical evaluation is needed to obtain these limiting values. For instance, using B=2 is more appropriate when $K>16$ and the highest modulation scheme is 64-QAM. In the case of 16-QAM, the number of users is reduced to $K>6$.

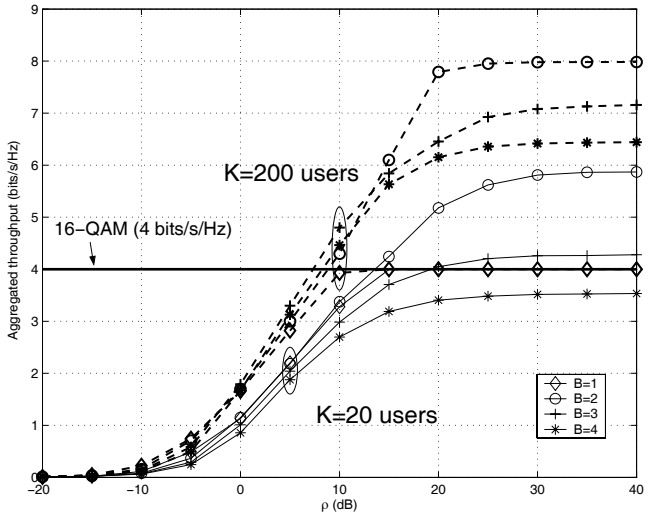

Fig. 2. Aggregated throughput vs. average SNR (ρ) for a different number of active beams (L=10 symbols). Horizontal line corresponds to the spectral efficiency of 16-QAM. Solid lines: K=20 users, dashed lines: K=200 users.

were increased[6].

A. Beam Selection Algorithms

In an ORB system, the *Optimum Beam Selection* algorithm consists in conducting an exhaustive search over all the possible subsets of beams and users. This procedure is computationally intensive as the number of SINRs to be tested grows exponentially with the number of available beams. More precisely, for a fixed number of active beams B, a total of $\binom{M}{B}KB$ SINRs must be computed in order to find the best transmission configuration. Then, by considering all the possible number of active beams, the number of SINR computations is increased to a total of $\sum_{B=1}^{M}\binom{M}{B}KB = KM2^{M-1}$ operations. Further, this search requires all the gains $|\mathbf{h}_k^T\mathbf{w}_i|^2$ to be known for any user-beam pair. Therefore, it is necessary for each user to report M real numbers to the BS over the feedback channel. It is worth noting that the amount of feedback required by this strategy is higher than the conventional ORB strategy (only one real number plus $\log_2 M$ bits for indicating the best beam) but, as shown in the next subsection, this amount of feedback increase results in significant gains in terms of sum-rate and aggregated throughput performance.

In order to reduce the computational complexity, we next present some sub-optimum approaches. Notice that the amount of feedback required by these strategies (except for the restricted beam selection strategy) is the same as that required by the optimum beam selection approach.

1) Bottom-up Beam Selection: The algorithm starts by selecting the best user for each beam in terms of the performance metric. After that, the users selected in the first step (with their associated beams) are grouped in all the possible subsets of two users in order to find the best combination with two active beams. The algorithm is iterated until the combination B=M active beams is reached. Basically, the objective is to reduce the computational cost by focusing on the users achieving the highest gains with only one active beam (i.e., in the absence of interference). By doing so, KM computations are still needed in the first level but only $\binom{M}{B}B$ operations for B=2, ..., M. As a result, complexity drops to $M2^{M-1} + M(K-1)$ SINR computations.

2) Top-down Beam Selection: In the bottom-up procedure, we restrict the search to the users maximizing system performance when only one beam is active. However, this subset of users may not be adequate when the number of beams increases and interference comes into play. For this reason, we propose a similar approach where the recursion is started by activating the maximum number of available beams (B=M). Then, an exhaustive search is performed with the aim of finding the best sub-set of users and beams in that configuration. After that, the algorithm is iterated in the reverse ordering by testing all the possible combinations with a lower number of active beams in each step. Again, user-beam pairs chosen in the first step are retained. One can easily verify that the number of SINR operations is equal to that of the bottom-up approach.

3) Greedy Beam Selection: The greedy beam selection procedure extend the search to a larger set of users. Specific details about the proposed greedy algorithm can be found in Table II but essentially it consists in selecting in each step the pair user-beam leading to a higher increase of sum-rate (or aggregated throughput). The algorithm is iterated until the configuration with all the active beams is reached and, then, the best subset with $B = j^*$ active beams is selected. The algorithm performs $(M - B + 1)K$ SINR computations in each iteration with a total computational cost of $\sum_{B=1}^{M}(M-B+1)K = \frac{KM}{2}(M+1)$ SINR operations.

4) Enhanced Greedy Beam Selection: In the greedy beam selection scheme proposed above, the overall performance depends on the user-beam pair of the first iteration (i.e., in the absence of inter-user interference). Instead, we can defer such decision to the second iteration where some inter-user interference is already present. In other words, we initialize the algorithm by identifying the *best* user for each beam $i = 1, ..., M$ (i.e., in the absence of interference). Then, we run the greedy algorithm M times taking as a starting point each user-beam pair obtained in the initialization. As a result, $\frac{KM}{2}(M+1) - K(M-1)$ operations should be done each time the algorithm is run and, since this procedure should be repeated M times, the total computational complexity amounts to $\frac{KM}{2}(M^2 - M + 2)$ SINR computations.

5) Restricted Beam Selection: Finally, we present a methodology where the optimum beam selection procedure is restricted to a predetermined number of active beams B. In other words, all the possible transmission configurations with B active beams are tested only. By doing so, the number of SINR computations drops to $\binom{M}{B}KB = \frac{KM!}{(M-B)!(B-1)!}$ operations. This strategy is very appropriate in those situations where the optimum number of active beams is known beforehand. Due to the complexity of the sum-rate and aggregated throughput expressions, obtaining the optimum number of beams (in terms of the SNR and K) is a quite complicated task

[6]In the single beam case (B=1), for instance, the number of equivalent users it is equal to MK. This is because $\mathsf{SINR}_{k,i}$ for $k = 1, ..., K$ and $i = 1, ..., M$ are i.i.d distributed in this case. For a higher number of active beams, however, the SINRs of the different beam sub-sets may be correlated and the diversity increase cannot be easily obtained.

TABLE II
GREEDY BEAM SELECTION ALGORITHM

1. Set j=1, $\mathcal{K}_1=\{1,\ldots,K\}$ and $\mathcal{B}_1=\{\mathbf{w}_1,\ldots,\mathbf{w}_M\}$.
2. Compute the best pair user-beam for the case with only one active beam as:
$$(k_1,i_1)=argmax_{(k,i)}\rho|\mathbf{h}_k^T\mathbf{w}_i|^2,\ \forall(k,i)\in\mathcal{K}_1\times\mathcal{B}_1$$
3. Compute $R_1=\Theta(\rho|\mathbf{h}_{k_1}^T\mathbf{w}_{i_1}|^2)$, where $\Theta(x)=\begin{cases}\log_2(1+x) & \text{for sum-rate}\\ b_{m_i}(1-\alpha_{m_i}e^{-\beta_{m_i}x})^L & \text{for aggregated throughput}\end{cases}$
4. Set $j=j+1$, $\mathcal{K}_j$=$\mathcal{K}_{j-1}-\{k_{j-1}\}$ and $\mathcal{B}_j$=$\mathcal{B}_{j-1}-\{\mathbf{w}_{i_{j-1}}\}$.
5. Compute the best pair user-beam that can be added to the system as:
$$(k_j,i_j)=$$
$$=argmax_{(k,i)}\left\{\Theta\left(\frac{|\mathbf{h}_k^T\mathbf{w}_i|^2}{j/\rho+\sum_{s=1}^{j-1}|\mathbf{h}_k^T\mathbf{w}_{i_s}|^2}\right)+\sum_{p=1}^{j-1}\Theta\left(\frac{|\mathbf{h}_{k_p}^T\mathbf{w}_{i_p}|^2}{j/\rho+|\mathbf{h}_{k_p}^T\mathbf{w}_i|^2+\sum_{\substack{s=1\\s\neq p}}^{j-1}|\mathbf{h}_{k_p}^T\mathbf{w}_{i_s}|^2}\right)\right\},$$
$$\forall(k,i)\in\mathcal{K}_j\times\mathcal{B}_j$$
6. Compute
$$R_j=\sum_{p=1}^{j}\Theta\left(\frac{|\mathbf{h}_{k_p}^T\mathbf{w}_{i_p}|^2}{j/\rho+\sum_{\substack{s=1\\s\neq p}}^{j}|\mathbf{h}_{k_p}^T\mathbf{w}_{i_s}|^2}\right)$$
7. If $j<M$, go to step 4. Otherwise go to step 8.
8. Set $j^*=argmax_jR_j$.
9. The algorithm is finished and the set of selected users and beams is the following:
$$(k_1,\mathbf{w}_{i_1}),...,(k_{j^*},\mathbf{w}_{i_{j^*}})$$

and an exhaustive search is required. Operating points can be obtained off-line and provided to the base station. However, by taking into consideration the analysis carried out in Sections IV and V, some rules can be derived. In terms of sum-rate performance, the optimum strategy is using a single active beam when the high-SNR scenario is considered, whereas the use of several active beams pays off for decreasing values of SNR[7]. When the aggregated throughput is considered, the single beam solution quickly saturates in the high-SNR region and (according to the size of the highest order modulation scheme) using a higher number of active beams could be a more appropriate choice. It is in the mid-SNR region where spatial multiplexing gains are efficiently exploited, since the impact of inter-beam interference on PER performance restricts its use in the low-SNR region. For both the sum-rate and aggregated throughput metrics, one can enlarge the SNR regions where adopting a multi-beam strategy is beneficial by increasing K.

Besides, since the number of active beams is fixed a priori, this strategy allows for a reduction in the amount of feedback. This is because each user can report the highest SINRs for each configuration instead of all the gains $|\mathbf{h}_k^T\mathbf{w}_i|^2$ (required by all the previous strategies) with a reasonable amount of feedback. Then, by sending only SINRs associated to a limited number of transmission configurations, sub-optimum approaches can be derived in terms of system performance vs. feedback requirements (further details are given in the next section). In the sequel, this algorithm will be called BSX, where X stands for the number of active beams.

[7]As previously commented, we consider $\rho=0$ dB as the low-SNR region due to practical constraints. For lower values of SNR, using only one active beam is the optimum strategy when a beam selection approach is considered [16].

B. Numerical Results and Discussion

We consider a system with a number of active users in the range $K=10,...,100$ transmitting data packets with L=10 symbols in each and with M=3 antennas at the base station. Notice that, in the proposed scenario, results obtained with the BS3 strategy will be equivalent to those obtained with ORB strategy with all the active beams (B=M=3). Then, these results can be considered in the analysis of the gain obtained with the proposed beam selection strategies with respect to the conventional ORB technique. Finally, it is also worth noting that both the restricted (BSX) and optimum beam selection (Optimum BS) algorithms presented in this work coincide with the static and dynamic ORBF/SBS strategies independently proposed in [16].

In Fig. 3, the different beam selection methodologies are compared in terms of sum-rate performance. We start by analyzing the low-SNR case (ρ=0dB). As expected, the best performance is obtained with the exhaustive search. Regarding the sub-optimum approaches, performance losses can be observed for both the bottom-up and greedy methodologies, whereas most of the sum-rate gains can be achieved with the top-down and enhanced greedy approaches. This is because using several active beams may be beneficial when the SNR is low and, then, incorrect decisions made in the first step of the greedy and bottom-up algorithms penalize system performance. This effect is even clearer when K increases. As for the restricted beam selection procedures, the best results are obtained with BS2. It is worth noting that better performance can be obtained with BS3 in scenarios with $M>5$ antennas. In those situations,

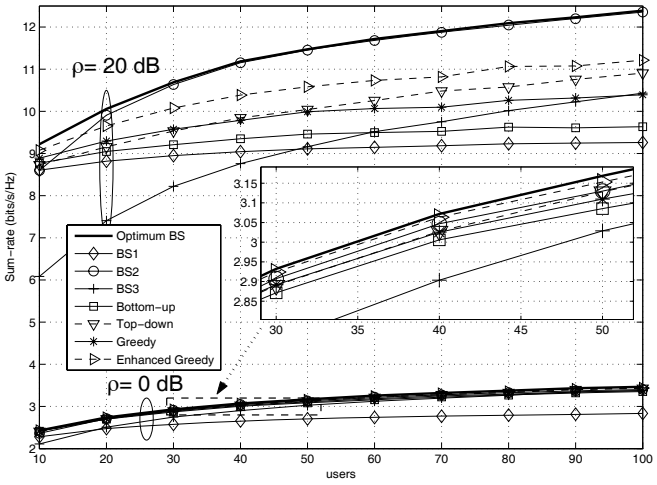

Fig. 3. Sum-rate vs. users for the different beam selection procedures. Top, ρ= 20 dB. Bottom, ρ= 0 dB.

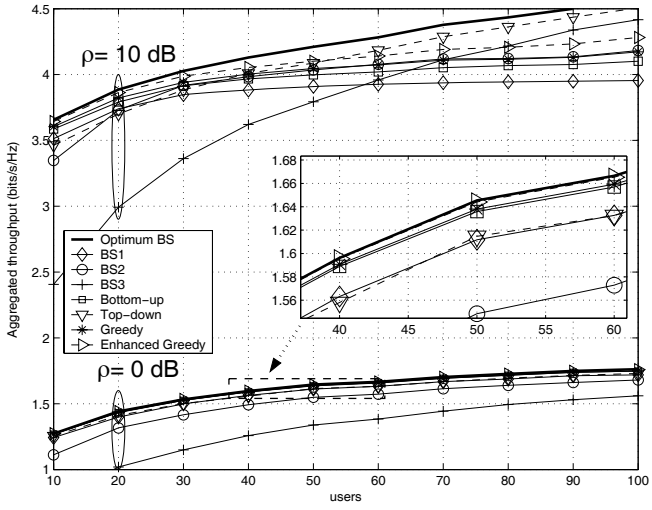

Fig. 4. Aggregated throughput vs. users for the different beam selection procedures (L=10 symbols). Top, ρ= 10 dB. Bottom, ρ= 0 dB.

however, the complexity of the system and the amount of information to be conveyed in the feedback channel becomes inappropriate for practical implementation and, for that reason, we have considered M=3.

When the SNR increases (see the top part of Fig. 3), the system becomes interference-limited. As a result, system behavior is more sensitive to the accuracy of the beam selection procedure. One can also observe that results associated to BS1 improve, whereas performance associated with BS3 deteriorates. This is because the optimum solution tends to use a reduced number of active beams when the SNR increases. This effect can be clearly observed in a high-SNR scenario, say ρ>30 dB (not shown here for brevity).

For the throughput case, however, a different behavior can be observed. As discussed in Section V, the most appropriate solution in low-SNR scenarios is to use only one active beam. Then, as confirmed by the results obtained for ρ=0 dB in Fig. 4, BS1 and those recursive techniques starting with one active beam (bottom-up, greedy and enhanced greedy) perform much

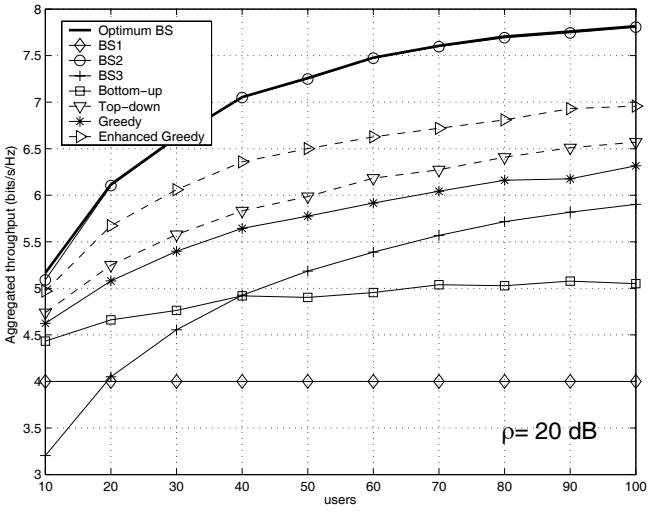

Fig. 5. Aggregated throughput vs. users for the different beam selection procedures (L=10 symbols, ρ= 20 dB).

TABLE III
COMPUTATIONAL COMPLEXITY FOR THE DIFFERENT APPROACHES IN TERMS OF SINR COMPUTATIONS (K=20 USERS).

	M=1	M=2	M=3	M=4	M=5
Optimum BS	20	80	240	640	1600
Bottom-up/Top-down	20	42	69	108	175
Greedy	20	60	120	200	300
Improved Greedy	20	80	240	560	1100
BS-1	20	40	60	80	100
BS-2	-	40	120	240	400
BS-3	-	-	60	240	600
BS-4	-	-	-	80	400
BS-5	-	-	-	-	100

better. However, throughput saturates when the SNR grows and adding more active beams pays off then (see Figs. 4 and 5). For moderate SNRs and a high number of users, better results are obtained with the top-down and BS3 approaches (see the high-users region of the ρ=10 dB case). Otherwise, BS2 seems to be a reasonable strategy due to its simplicity and overall performance.

Finally, in Table III we compare the different schemes in terms of computational complexity. The sub-optimum ones, either recursive or not, lead to substantial savings with respect to conducting an exhaustive search, which compensate for their performance losses in some cases. Out of the sub-optimum approaches, enhanced greedy is the one exhibiting the highest complexity.

VII. IMPACT OF FEEDBACK QUANTIZATION ON AGGREGATED THROUGHPUT

In this section, we focus on the case where the SINRs conveyed over the feedback channel are quantized versions of the analog ones. Motivated by the work in [8], this section is devoted to find out how many feedback bits are required to exploit MUD in an ORB context. Finally, a beam selection approach trading-off system performance vs. feedback requirements is proposed.

A. Aggregated Throughput with Quantized Feedback

Let $\mathcal{Q}=\{q_1, q_2, ..., q_{2^{L_q}}\}$ be the set of quantization levels. After an arbitrary user k identifies the beam i^* with the highest SINR, $\gamma_{k,i^*}=\max_{i\in\mathcal{B}} \mathsf{SINR}_{k,i}$, it is quantized according to the rule "$Q(\gamma_{k,i^*})=\gamma_{q_j}$ if $\gamma_{q_j}\leq\gamma_{k,i^*}<\gamma_{q_{j+1}}$", where γ_{q_j} $(j=1, ..., \text{card}(\mathcal{Q}))$ are the different SINR thresholds associated with the quantization levels. Notice that such quantization rule results into a conservative, but reliable, assignment of AMC modes. Next, an $L_t=L_q+L_b$-bit message is sent over the feedback channel, with $L_b=\lceil log_2(B)\rceil$ bits dedicated to identify the selected beam.

Now, we define $\mathcal{A}_{k,i}$ as the event that user k is selected on beam i, and we obtain the probability of $\mathcal{A}_{k,i}$ conditioned on the fact that $\gamma_{k,i}$ belongs to the level q_j as [8]:

$$\text{Prob}(\mathcal{A}_{k,i}|\gamma_{k,i}\in q_j) = \sum_{l=0}^{K-1}\frac{1}{l+1}\binom{K-1}{l}\text{Prob}\,(\#\text{users different from } k \text{ in } q_j=l) \times \text{Prob}(\#\text{users different from } k \text{ in levels lower than } q_j=K-l-1)$$

Each term in the summation takes into consideration that, apart from user k, l users lie in quantization level q_j and that the rest of users must lie in lower quantization levels. The term $1/(l+1)$, on the other hand, is referred to the probability that user k is scheduled for transmission for each value of l, since a random user must be selected if more than one user lie in q_j.

In an homogeneous scenario, $\text{Prob}(\mathcal{A}_{k,i}|\gamma_{k,i}\in q_j)$ does not depend on k or i. More specifically, by modeling the SINRs of all the users with a generic random variable γ with CDF and pdf expressions given by Eqs. (4) and (5), respectively, the following expression results:

$$\text{Prob}(\mathcal{A}_{k,i}|\gamma_{k,i}\in q_j) = \sum_{l=0}^{K-1}\frac{1}{l+1}\binom{K-1}{l}\left(\text{Prob}(\gamma\in q_j)\right)^l \times\left(\text{Prob}\left(\gamma\in\bigcup_{p<j} q_p\right)\right)^{K-l-1}$$

After some algebraic manipulation and bearing in mind that $\text{Prob}(\gamma\in q_j) = F_{\mathsf{SINR}}(\gamma_{q_{j+1}}) - F_{\mathsf{SINR}}(\gamma_{q_j})$ and $\text{Prob}\left(\gamma\in\bigcup_{p<j} q_p\right) = F_{\mathsf{SINR}}(\gamma_{q_j})$, it can be readily shown that

$$\text{Prob}(\mathcal{A}_{k,i}|\gamma_{k,i}\in q_j) = \frac{1}{K}\frac{\left(F_{\mathsf{SINR}}(\gamma_{q_{j+1}})\right)^K - \left(F_{\mathsf{SINR}}(\gamma_{q_j})\right)^K}{F_{\mathsf{SINR}}(\gamma_{q_{j+1}}) - F_{\mathsf{SINR}}(\gamma_{q_j})} \tag{17}$$

The throughput share corresponding to user k on beam i follows as:

$$\eta_{k,i} \simeq \sum_{j=1}^{\text{card}(\mathcal{Q})}\text{Prob}(\mathcal{A}_{k,i}|\gamma_{k,i}\in q_j) b_{m_j} \times\int_{\gamma=\gamma_{q_j}}^{\gamma_{q_{j+1}}}(1-\alpha_{m_j}e^{-\beta_{m_j}\gamma})^L f_{\mathsf{SINR}}(\gamma)d\gamma\ , \tag{18}$$

where each integration interval corresponds to the expected throughput in each quantization level. Besides, we assume that the modulation scheme for quantization level q_j is selected according to the quantized value of $\gamma_{k,i}$, that is $m_j = m \iff \gamma_{th,m}\leq\gamma_{q_j}<\gamma_{th,m+1}$.

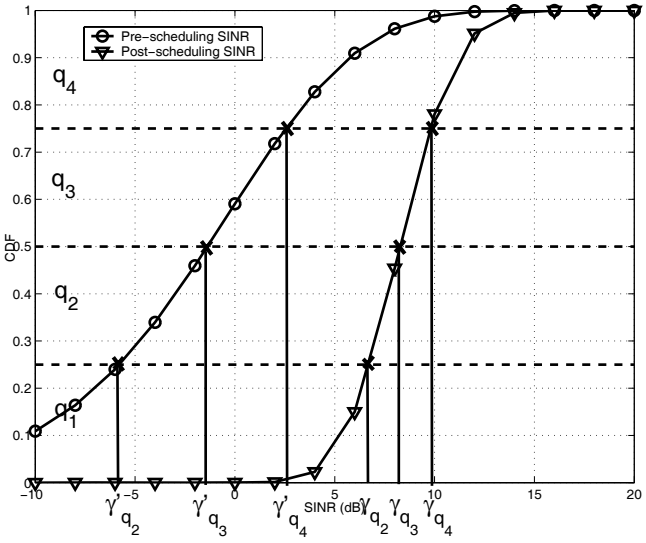

Fig. 6. Pre- and post-scheduling quantization procedures for the SINR quantization (ρ=10 dB, K=20 users, B=2 active beams).

Finally, by plugging Eqs. (17) and (5) into Eq. (18) and summing up over users (k=1, ..., K) and active beams ($i\in\mathcal{B}$), we can write the aggregated throughput expression as:

$$\eta(\mathcal{B},\mathcal{Q}) \simeq \sum_{k=1}^{K}\sum_{i\in\mathcal{B}}\eta_{k,i} = B\sum_{j=1}^{\text{card}(\mathcal{Q})} b_{m_j} \times\sum_{l=0}^{L}\binom{L}{l}(-\alpha_{m_j})^l\frac{\left(F_{\mathsf{SINR}}(\gamma_{q_{j+1}})\right)^K - \left(F_{\mathsf{SINR}}(\gamma_{q_j})\right)^K}{F_{\mathsf{SINR}}(\gamma_{q_{j+1}}) - F_{\mathsf{SINR}}(\gamma_{q_j})} \times\int_{\gamma=\gamma_{q_j}}^{\gamma_{q_{j+1}}}\left(\frac{B}{\rho}(1+\gamma)+B-1\right)\frac{e^{-\gamma\left(\beta_{m_j}l+\frac{B}{\rho}\right)}}{(1+\gamma)^B}d\gamma\ .$$

The integral term in the above expression resembles that of Eq. (21) in Appendix. Therefore, one can once again follow the same procedure to finally obtain:

$$\eta(\mathcal{B},\mathcal{Q}) \simeq B\sum_{j=1}^{\text{card}(\mathcal{Q})} b_{m_j}\sum_{l=0}^{L}\binom{L}{l}(-\alpha_{m_j})^l \times\frac{\left(F_{\mathsf{SINR}}(\gamma_{q_{j+1}})\right)^K - \left(F_{\mathsf{SINR}}(\gamma_{q_j})\right)^K}{F_{\mathsf{SINR}}(\gamma_{q_{j+1}}) - F_{\mathsf{SINR}}(\gamma_{q_j})}e^{\mu_q}\mu_q^{c_q} \times\left[\frac{B}{\rho\mu_q}\left(\Gamma\left(1-c_q,(1+\gamma_{q_j})\mu_q\right)-\Gamma\left(1-c_q,(1+\gamma_{q_{j+1}})\mu_q\right)\right) +(B-1)\left(\Gamma\left(-c_q,(1+\gamma_{q_j})\mu_q\right)-\Gamma\left(-c_q,(1+\gamma_{q_{j+1}})\mu_q\right)\right)\right] \tag{19}$$

where $\mu_q=\beta_{m_j}l+\frac{B}{\rho}$ and $c_q=B-1$.

B. Quantization Law

So far, the quantization law has been left as a free parameter. Obtaining the set of quantization thresholds maximizing the throughput expression (19) is quite involved. However, we can simplify this issue by recalling that, according to the

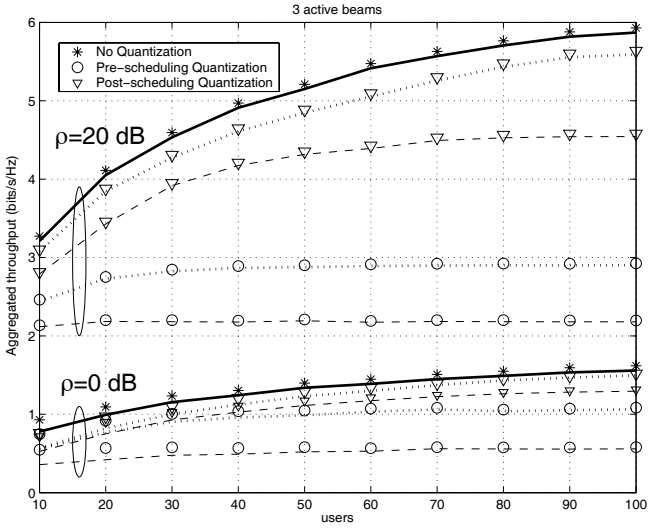

Fig. 7. Aggregated throughput vs. number of users for the different quantization methodologies and number of quantization bits, L_q (L=10 symbols, B=3 active beams, dotted lines: 4 quantization bits, dashed lines: 2 quantization bits, symbols: simulated results, curves: analytical expressions).

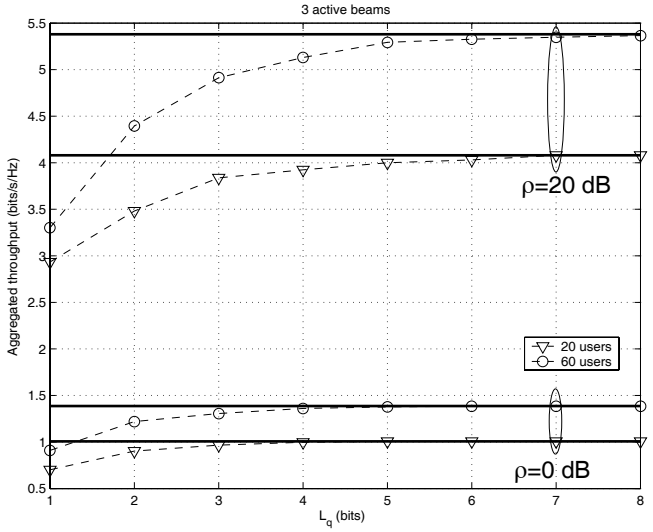

Fig. 8. Aggregated throughput vs. number of quantization bits (L_q) for different number of users (L=10 symbols, B=3 active beams, solid lines: unquantized throughput).

adopted scheduling rule, only the highest SINRs are relevant to take advantage of multi-user gains. Therefore, we expect that throughput maximization is obtained from a non-uniform quantization law with smaller quantization intervals in the high-SINR region. As depicted in Fig. 6, the quantization thresholds could be given by the inverse of either the *post-scheduling* CDF function or the *pre-scheduling*, i.e. individual, CDFs. Of course, a quantization law tailored to the post-scheduling SINR is expected to give better results since this is directly related with the scheduling rule. Numerical evaluation in Subsection VII-C confirms this conjecture. In summary, the SINR thresholds related to the different quantization levels are selected as:

$$\gamma_{q_j} = F^{-1}_{\mathsf{SINR}^*}\left(\frac{j-1}{2^{L_q}}\right) \text{ for } j = 1, ..., \text{card}(\mathcal{Q})$$

with γ_{q_1}=0 and $\gamma_{q_{\text{card}(\mathcal{Q})+1}}$=$\infty$. Notice this is a extension of the quantization procedure proposed in [26] in the sense that a multi-beam scenario is considered and a *cross-layer* strategy is adopted. The last affirmation comes from the fact that quantization levels in the physical layer depend on the number of admitted users decided by the access control mechanism in the data link layer. As a final remark, it is worth pointing out that the proposed method may have practical limitations as all the users need the value of K. Then, the BS must broadcast this information to the users periodically. As a result, the overall throughput of the system may be affected (specially in scenarios with bursty traffic) but the analysis of such effect is out of the scope of this work.

C. Numerical Results and Discussion

As in the previous section, we restrict the analysis to a scenario with M=3 transmit antennas. First, we evaluate in Fig. 7 performance in terms of aggregated throughput for the quantization methodologies based on pre- and post-scheduling statistics. The post-scheduling based criterion significantly outperforms its pre-scheduling counterpart for the whole range

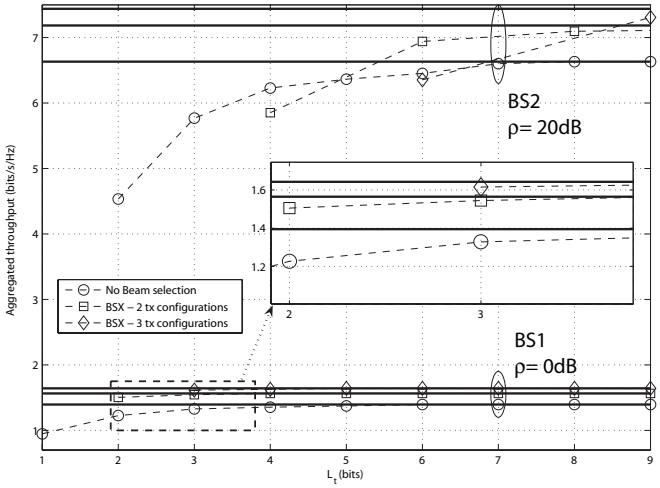

Fig. 9. Aggregated throughput vs. number of feedback bits (L_t) for different restrictions in the BSX procedure (Solid lines: unquantized throughput, L=10 symbols, K=60 users). Top, BS2 in a scenario with ρ=20 dB. Bottom, BS1 in a scenario with ρ=0 dB.

of users. For an increasing number of active users, the performance gap gets larger since the increased clipping rate (i.e., the SINR of the scheduler user is potentially higher than the highest quantization threshold) penalizes much more the quantization law based on pre-scheduling statistics. When adopting a quantization law based on the post-scheduling CDF most of the MUD gain can be efficiently captured with L_q=4 or L_q=2 bits. The curves depicted in Fig. 8 provide some more insight into the impact of quantization on system performance. First, the higher the SNR the larger the impact of quantization since, in this case, the range of SINR fluctuations is larger. As commented above, this is emphasized in scenarios with a high number of users. However, most of the MUD can be effectively captured by a very low number of bits. More precisely, for the worst case (ρ=20 dB and K=60 users), the proposed quantization law attains approximately 81% and 92% of the *analog* throughput by just using 2 or 3 bits,

respectively.

Finally, we focus on the BS1 and BS2 procedures with the aim of analyzing their performance as a function of the total number of feedback bits L_t. As shown in Section VI, these schemes exhibit remarkable performance vs. complexity trade-off and allow the derivation of sub-optimum approaches according to the restrictions in the feedback channel. For instance, consider a scenario with M=3 available beams. In this case, three transmission configurations are available for both the BS1 and BS2 procedures. That is, with B=2 the beam subsets available at the base station are $\mathcal{B}_1$=$\{1,2\}$, $\mathcal{B}_2$=$\{1,3\}$ and $\mathcal{B}_3$=$\{2,3\}$; whereas for the single beam case the subsets are $\mathcal{B}_1$=$\{1\}$, $\mathcal{B}_2$=$\{2\}$ and $\mathcal{B}_3$=$\{3\}$. The system can then be simplified by considering only one (No Beam Selection, only $\mathcal{B}_1$ is available) or two transmission configurations in the selection procedure (BSX-2 tx, $\mathcal{B}_1$ and $\mathcal{B}_2$ are available). By doing so, the total number of bits L_t required in the feedback channel is lower. For the BS1 case, only the quantized version of the highest SINR and the beam index must be sent to the BS. As a consequence, the number of required bits amounts to $L_t = L_q + \lceil \log_2 N_{tx_{conf}} \rceil$, where $N_{tx_{conf}}$ stands for the number of transmission configurations. Conversely, $L_t = N_{tx_{conf}}(L_q + \lceil \log_2 B \rceil)$ bits are necessary in the BS2 scheme since both the quantized versions of the SINRs and the index beams must be reported.

First, we analyze BS1 in the low-SNR regime (Fig. 9, bottom). As long as we increase the number of quantization bits, performance improves. Further, if some of those bits are dedicated to incorporate beam selection capabilities, larger gains can be achieved. As an example, by using L_t=2 bits we gain 29.46% with respect to the L_t=1-bit case if both bits are used for quantization, whereas the gain rises to 58.34% if one of those bits is used for beam selection (BS1-2 tx). For the L_t = 3-bits case, the gain with BS1-3 tx amounts to 69.96%. As for the high-SNR region (Fig. 9, top), similar conclusions follow to BS2. In particular, a good performance vs. feedback trade-off is provided by BS2-2 tx as substantial gains can be obtained with only L_t=6 bits.

VIII. Conclusions

In this paper, we have investigated the performance of ORB in a context where the *number* of active beams is used as design parameter and a *realistic* number of users has been considered.

From the analysis in terms of sum-rate we have proved that in interference-limited scenarios one single active beam turns out to be the best strategy whereas a higher number of beams is preferred in a noise-limited context. The throughput analysis (where granularity and saturation effects result from the limited number of AMC modes) has shown that in the interference-limited scenario, the single-beam solution quickly saturates for small constellation sizes and one should rather use B=2 beams. In noise-limited scenarios, the (high) PERs experienced by the smallest constellation size often make multiple-beam solutions barely recommended.

On the basis of the conclusions above, we have proposed optimum and sub-optimum beam selection schemes. Their sum-rate or throughput performance has been assessed thus proving that *restricted beam selection* exhibits a good performance-complexity trade-off.

Finally, we have analyzed the impact of SINR quantization on system performance, showing that 2 quantization bits is enough to have the MUD gain offered by ORB. Afterwards, we have proposed a beam selection algorithm capable of trading-off system performance vs. feedback requirements. In particular, we have shown that ORB performance can be considerably improved by requiring a few bits in the feedback channel with the help of *restricted beam selection*.

Appendix

In order to derive a closed-from expression of the aggregated throughput, one should solve expression (15). By plugging (6) into (15) and using the binomial expansion, the following expression results:

$$\eta(\mathcal{B}) = B \sum_{m=1}^{\mathrm{card}(\mathcal{M})} b_m K \sum_{l=0}^{L} \binom{L}{l} (-\alpha_m)^l \sum_{k=0}^{K-1} \binom{K-1}{k} (-1)^k \times \int_{\gamma=\gamma_{th,m}}^{\gamma_{th,m+1}} \left(\frac{B}{\rho}(1+\gamma) + B - 1 \right) \frac{e^{-\gamma\left(\beta_m l + \frac{B}{\rho}(k+1)\right)}}{(1+\gamma)^{k(B-1)+B}} d\gamma \tag{20}$$

where it is observed that the problem is reduced to solve an integral of the type:

$$\int_{t=u}^{v} \frac{e^{-at}}{(1+t)^n} dt = \int_{t=u}^{\infty} \frac{e^{-at}}{(1+t)^n} dt - \int_{t=v}^{\infty} \frac{e^{-at}}{(1+t)^n} dt \quad a>0; n=1,2,... \tag{21}$$

By using the change of variables $x = (1+t)a$, one can re-writte anyone of the integrals above as $\int_{t=u}^{\infty} \frac{e^{-at}}{(1+t)^n} dt = e^a a^{n-1} \int_{x=(1+u)a}^{\infty} e^{-x} x^{-n} dx$ and notice that can be easily solved by resorting to the identity [22, Eq. 8.350.2]:

$$\int_{t=u}^{\infty} \frac{e^{-at}}{(1+t)^n} dt = e^a a^{n-1} \Gamma(1-n, (1+u)a) \tag{22}$$

where $\Gamma(\alpha, x)$ stands for the complementary incomplete gamma function ($\Gamma(\alpha, x) \triangleq \int_x^{\infty} e^{-t} t^{\alpha-1} dt$). Finally, by using Eq. (22) in Eq. (21) one can solve Eq. (20) and verify that Eq. (16) holds.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments on the paper.

References

[1] D. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," in *Proc. IEEE ISIT*, 1997.

[2] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE ICC*, 1995.

[3] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian MIMO broadcast channel," in *Proc. CISS*, Princeton, USA, 2004.

[4] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.

[5] G. Dimic and N. D. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Trans. Signal Processing*, vol. 53, no. 10, pp. 3857–3868, Oct. 2005.
[6] T. Yoo and A. Goldsmith, "On the optimality of multi-antenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Select. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.
[7] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channel with partial side information," *IEEE Trans. Inform. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.
[8] F. Floren, O. Edfors, and B. Molin, "The effect of feedback quantization on the throughput of a multiuser diversity scheme," in *Proc. Globecom*, 2003.
[9] S. Sanayei and A. Nosratinia, "Exploiting multiuser diversity with only 1-bit feedback," in *Proc. IEEE WCNC*, 2005.
[10] M. Kountouris and D. Gesbert, "Memory based opportunistic multi-user beamforming," in *Proc. IEEE ISIT*, 2005.
[11] ——, "Robust multi-user opportunistic beamforming for sparse networks," in *Proc. IEEE SPAWC*, 2005.
[12] M. Kobayashi, G. Caire, and D. Gesbert, "Opportunistic beamforming vs. space-time coding in a queued downlink," in *Proc. IST Summit*, Dresden, Germany, 2005.
[13] J. L. Vicario, R. Bosisio, U. Spagnolini, and C. Anton-Haro, "A throughput analysis for opportunistic beamforming with quantized feedback," in *Proc. IEEE PIMRC*, Sept. 2006.
[14] ——, "Adaptive beam selection techniques for opportunistic beamforming," in *Proc. IEEE PIMRC*, Sept. 2006.
[15] K. Zhang and Z. Niu, "Random beamforming with multi-beam selection for MIMO broadcast channels," in *Proc. IEEE ICC*, June 2006.
[16] J. Wagner, Y. Liang, and R. Zhang, "Random beamforming with systematic beam selection," in *Proc. IEEE PIMRC*, Sept. 2006.
[17] J. Diaz, O. Simeone, and Y. Bar-Ness, "Sum-rate of MIMO broadcast channels with one bit feedback," in *Proc. IEEE ISIT*, 2006.
[18] T. Marzetta and B. Hochwald, "Capacity of a mobile multiple-antenna communication link in rayeligh flat fading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 139–157, Jan. 1999.
[19] R. Bosisio, J. L. Vicario, C. Anton-Haro, and U. Spagnolini, "Diversity-multiplexing tradeoff in multi-user scenario with selective feedback," in *Proc. IST Mobile & Wireless Communications Summit*, 2006.
[20] E. G. Larsson, "On the combination of spatial diversity and multiuser diversity," *IEEE Commun. Lett.*, vol. 8, no. 8, pp. 517–519, Aug. 2004.
[21] J. L. Vicario, "Antenna selection techniques in single- and multi-user systems: a cross-layer approach," Ph.D. dissertation, UPC, available for download at http://tes.uab.es/TSC/research/communications/publilopezv.htm, Sept. 2006.
[22] I. Gradsthteyn and I. Ryzhik, *Tables of Integrals; Series and Products*. New York: Academic, 1965.
[23] Q. Liu, S. Zhou, and G. Giannakis, "Cross-layer combinig of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746–1755, Sept. 2004.
[24] S. Chung and A. Goldsmith, "Degrees of freedom in adaptive modulation: a unified view," *IEEE Trans. Commun.*, vol. 49, no. 9, pp. 1561–1571, Sept. 2001.
[25] T. Starr, J. Cioffi, and P. Silverman, *Understanding Digital Subscriber Line Technology*. Prentice Hall, 1999.
[26] S. Sanayei and A. Nosratinia, "Opportunistic beamforming with limited feedback," in *Proc. Asilomar Conference on Signals, Systems and Computers*, 2005.

Jose Lopez Vicario (S'04-M'08) was born in Blanes, Spain, in 1979. He received both the degree in electrical engineering and the Ph.D. degree from the Universitat Politecnica de Catalunya (UPC), Barcelona, in 2002 and 2006, respectively. During 2002, he served as DSP programmer and participated in the Spanish government R&D project TIC99-0849. From October 2002 to September 2006, he was a Ph.D. candidate at UPC's Signal Theory and Communications Department and, from January 2003, he pursued his thesis at the Centre Tecnologic de Telecomunicacions de Catalunya (CTTC), where he was involved in several European R&D projects (IST ACE and IST NEWCOM). Since September 2006, he is an Assistant Professor at the Universitat Autonoma de Barcelona (UAB) teaching courses in digital communications and information theory. His research interests are in the area of cross-layer designs, MIMO systems, cooperative communications and network coding.

Roberto Bosisio received the Msc Degree (with honors) in Telecommunication Engineering from Politecnico di Milano in 2003. From March to September 2004 he worked as a visiting researcher at the Center for Signal Processing and Communication Research at the New Jersey Institute of Technology, Newark, USA. After this experience, Roberto pursued the PhD in Telecommunication Engineering and he received the Ph.D. degree in 2007 from Politecnico di Milano. His current research interest lies in the field of digital signal processing for digital communication with emphasys on multicarrier and multiantenna systems.

Carles Antón-Haro (M'99-SM'03) received his Ph.D. degree in Telecommunications engineering from the Technical University of Catalonia in 1998 (cum-laude). In 1999, he joined Ericsson Spain, where he participated in rollout projects of 2G and 3G mobile networks. Currently, he is with the Centre Tecnologic de Telecomunicacions de Catalunya (CTTC) as a Director of International Relations. His research interests are in the area of signal processing and digital communications, this including cross-layer designs, MIMO systems, opportunistic communications, wireless sensor networks, array signal processing for mobile communications, and wireless sensor networks. He has published +10 technical papers in IEEE journals as well as over 45 papers in international and national conferences. Dr. Anton-Haro is a Senior Member of the IEEE.

Umberto Spagnolini (SM'03) received the Dott. Ing. Elettronica degree (cum laude) from Politecnico di Milano, Milan, Italy, in 1988. Since 1988, he has been with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, where he is Full Professor in Telecommunications. He is the co-founder of WiSyTech (Wireless System Technology), a spinoff company of Politecnico di Milano on Software Defined Radio. His interests are in the area of statistical signal processing. The specific areas of interest include channel estimation and space-time processing for wireless communication systems; synchronization and cooperation in wireless sensor networks; parameter estimation/tracking and wavefield interpolation applied to UWB radar, oil exploration, and remote sensing. Dr. Spagnolini served on the editorial board and technical program committees of several conferences in all the areas of interests.