Optimal Resource Allocation in Relay-Assisted Cellular Networks With Partial CSI

Eduard Calvo, *Student Member, IEEE,* Josep Vidal, *Member, IEEE,* and
Javier Rodríguez Fonollosa, *Senior Member, IEEE*

Abstract—Emerging cellular networks are likely to handle users with heterogeneous quality of service requirements attending to the nature of their underlying service application, the quality of their wireless equipment, or even their contract terms. While sharing the same physical resources (power, bandwidth, transmission time), the utility they get from using them may be very different and arbitrage is needed to optimize the global operation of the network. In this respect, resource allocation strategies maximizing network utility under practical constraints are investigated in this paper. In particular, we focus on a cellular network with half-duplex, MIMO terminals and relaying infrastructure in the form of fixed and dedicated relay stations. Whereas orthogonal-frequency-division multiple access is assumed, it is seen as a frequency diversity enabler since path loss is the only channel state information (CSI) known at the transmitters, which is refreshed periodically. With this setup, the performance of a state-of-the art relay-assisted transmission protocol is characterized in terms of the ergodic achievable rates, for which novel concave lower bounds are developed. The use of these bounds allows us to derive two efficient algorithms computing resource allocations in polynomial time, which address the optimization of the uplink and downlink directions jointly. First, a global optimization algorithm providing one Pareto optimal solution maximizing network utility during the validity period of one CSI is studied, which acts as a performance upper bound. Second, a sequential optimization algorithm maximizing network utility frame by frame is considered as a simpler alternative. The performance of both schemes has been compared in practical scenarios, giving special attention to the performance-complexity and throughput-fairness tradeoffs.

Index Terms—Network utility maximization, optimization methods, resource allocation, wireless networks.

I. INTRODUCTION

A. Motivation

THE deployment of cellular networks has been traditionally associated to the provision of voice (and low-rate data) service to mobile users. The exclusivity of this purpose,

however, is in conflict with the ubiquitous availability of wireless equipment and the steadily increasing traffic demands arising from new interactive, multimedia services, which have opened the door to a plethora of new potential network scenarios. From interactive gaming to wireless broadband access, different services with heterogeneous quality of service (QoS) requirements shall converge to the same service network. Regarding this paradigm, we identify three central issues in prospective network design which motivate this paper:

- how to characterize the user experience of the different services of the network using homogeneous performance measures;
- how to dynamically arbitrate on the shared use of the limited transmission resources of the network by competing flows which are of different nature;
- how to extract the largest possible system spectral efficiency from the physical layer.

With this in mind, the optimization of the operation of the network is hence a matter of allocating resources (power, bandwidth, rate, transmission time) efficiently for uplink (UL) and downlink (DL) scheduled transmissions among the serving users such that some networkwide cost function involving their service experience is maximized along time.

B. Adopted Network Setup

In this paper, we tackle the network design problem adopting a cell-by-cell approach. Hence, we focus on a cell consisting of one base station (BS) serving M mobile stations (MSs, or users). To enable the realization of high spectral efficiencies and boost network performance, we assume that R relay stations (RSs) are deployed within the cell coverage area to enhance the communication between the users and the BS [1]–[5]. Interpreting the presence of relays as an extension of the network infrastructure enabling relay-assisted transmission, their locations are assumed to remain fixed, although they can indeed be optimized beforehand. All the terminals are assumed half-duplex for practical reasons.

Since the capacity of the relay channel is still an open problem (so is determining the optimal relaying strategy) we shall adopt here the cooperation protocol of [6, Prop. 2], based on the decode-and-forward strategy [7], which comprises essentially some of the protocols in [3], [5], [8], [9] as particular cases and is able to work with partial knowledge of the channel state. To make our approach more general, we let the BS, the RSs, and the MSs be equipped with an arbitrary number of antennas, denoted by n_{BS}, n_{RS}, and $n_{MS,m}$, respectively ($n_{MS,m}$ is the number of antennas of the mth MS, $1 \leq m \leq$ M) such that extra performance gains arising from MIMO [10]–[12] can be also captured.

Pursuing the application of our results to practical scenarios we are led to two important choices, the first one being the adoption

of orthogonal frequency division multiple-access (OFDMA). OFDMA can be efficiently implemented via FFT/IFFT, and it is able to combat the inherent frequency selectivity of wireless channels while at the same time allowing a modular tone-based multiplexing of users. In addition, it improves upon TDMA with respect to achievable rates and data latency, and allows for finer granularity in resource allocation [13], a must in wideband systems. For these and other reasons, it results appealing for upcoming wireless networking standards [14]–[16].

The second choice is related to the availability and quality of channel state information (CSI) at each network location (BS, RSs, and MSs). In relayless OFDMA networks, centralized perfect CSI of all the links (in the form of per-tone fading state knowledge) can be used to allocate resources adaptively. Hence, bandwidth, power, and rate can be optimally assigned to align with the instantaneous network conditions [17], yielding enormous performance gains. However, such perfect CSI is likely not to be available in all the $(R+M+RM)$ links of our scenario. On the one hand, the amount of processing required to take advantage of perfect CSI can be formidable (the complexity has been shown to be NP-hard even for a relayless network [18]) and possibly non-affordable. On the other hand, for sufficiently fast time-varying channels, the necessary CSI refresh interval can happen to exceed the capacity of the limited-rate feedback channels of the network. Even worse, propagation and processing delays on the feedback channels may result in outdated, useless CSI at the beginning of a resource allocation phase. Thus, unlike other works [17]–[21] we shall study the network scenario of all transmitters having perfect knowledge of the *path loss* of each of the channels, a slowly varying scalar parameter, but being ignorant of each per-tone fading state. Although explicit path loss estimation techniques are out of the scope of this paper, its accurate estimation seems reasonable provided that some pilot tones are placed within the transmission bandwidth, which is a common practice in OFDM-based standards such as the IEEE 802.16 suite [14]–[16], the 3GPP LTE [22], and WiMAX [23] for synchronization purposes.

With this setup, we aim at optimizing the network operation for maximizing network utility [21], [24]–[27] in a cell-by-cell approach. Centralized optimization is hence performed at each BS which, upon collection of CSI, takes scheduling decisions and implements resource allocation strategies shaping the instantaneous rates of all the users involved in its cell. One nice feature of our network operation design framework is that the network resources (time, frequency, power, and rate) devoted to UL and DL transmissions are optimized *jointly*, instead of allocating a given portion of total resources to each direction in each transmission frame and optimizing them separately.

C. Summary of Contributions

This paper proposes a centralized optimization framework for the maximization of the cell performance based on the user experience of each serving MS. Under the setup of Section I-B, the

---

CSI of all the links (path loss) is collected at the BS which, together with the QoS requirements of each UL and DL flow[1] and its current degree of fulfillment, decides the resource allocation strategy to be followed during some period of time. This strategy is based on the maximization of network utility, a cellwide performance measure which combines the service satisfaction of all the users, and has given rise to the following contributions to the problems raised in Section I-A.

- User satisfaction is measured using utility functions. Thus, the same network infrastructure can flexibly reconfigure to optimally serve a variety of scenarios by properly choosing the user utility function of each service under operation such that their different profiles are conveniently reflected.
- An algorithm to efficiently compute a global optimal resource allocation strategy in polynomial time (by solving a series of convex optimization problems) is proposed. It is benchmarked against other simpler, suboptimal strategies able to retain a large fraction of performance with significant complexity savings.
- The optimal operation of the network that maximizes network utility is essentially cross-layer, as the joint optimization of user scheduling, resource allocation, and relay-assisted transmission is involved for UL and DL directions.
- In characterizing the performance of the adopted relay-assisted transmission protocol, tight concave lower bounds to the ergodic capacity of MIMO and distributed MIMO channels are obtained which may find applications outside the scope of this paper.

D. Outline of the Paper

This paper is structured as follows. Section II describes the adopted transmission strategy for OFDMA with partial CSI and some preliminaries regarding key system parameters. Next, Section III addresses the transmission protocol for relay-assisted communication. Its cellwide short term performance is analytically characterized in Section IV in terms of instantaneous achievable rate regions. Then, Section V builds upon this to i) introduce user utility functions as a useful tool to characterize user satisfaction with services of different nature, ii) pose optimal network strategy as the solution to an optimization problem which aims at maximizing network utility, and iii) propose an iterative algorithm to compute a global optimal solution to this problem in polynomial time. In addition, a reduced-complexity algorithm computing a suboptimal network strategy is also proposed and benchmarked against the global optimal in Section VI, where simulation results of practical scenarios are provided. Finally, Section VII concludes the paper summarizing results and sketching lines for future work.

E. Notation

Throughout this paper, boldface lower-case letters denote column vectors, with $\mathbf{0}_n$ and $\mathbf{1}_n$ standing for the all-zero and all-one column vectors of length n, respectively. We shall denote by x_i the ith entry of vector $\mathbf{x}$, and use $\geq$ and $\leq$ for scalar and component-wise inequalities indistinctly. Similarly, $\mathbf{z} = \min\{\mathbf{x}, \mathbf{y}\}$ denotes the vertical stacking of the component-wise minimum of two vectors. Boldface upper-case letters are used for matrices, with $\mathbf{I}_n$ standing for the $n \times n$ identity matrix and $A_{i,j}$ denoting the entry of the ith row and

jth column of matrix $\mathbf{A}$, whose transpose and Hermitian are $\mathbf{A}^T$ and $\mathbf{A}^\dagger$, respectively. The ith ordered eigenvalue (singular value) of an square (arbitrary) matrix $\mathbf{A}$ is denoted by $\lambda_i(\mathbf{A})$, where $\lambda_i(\mathbf{A}) \geq \lambda_{i+1}(\mathbf{A})$. Whenever needed, the superscript $(\cdot)^\star$ shall denote the optimal value of a variable.

II. SYSTEM MODEL AND PRELIMINARIES

Consider the network setup described in Section I-B, where the BS, the RSs, and the MSs are power constrained to $p_{\mathrm{BS}}^{\max}$, $p_{\mathrm{RS}}^{\max}$, and $p_{\mathrm{MS}}^{\max}$, respectively. In every transmission frame interval, denoted by T, the same network bandwidth B is used in the UL and DL phases, of adjustable duration via TDD.[2] In each of them, the communication of each BS–MS pair is assisted by one RS. Let us denote the RS attached to the mth MS by $\mathrm{RS}(m) \in \{1, \ldots, \mathrm{R}\}$. The RS assignment of the network is hence described by the connectivity matrix $\mathbf{L} \in \{0,1\}^{\mathrm{R} \times \mathrm{M}}$, where $L_{r,m} = \delta[r - \mathrm{RS}(m)]$ and $\delta[\cdot]$ is the Kronecker delta. Note that each BS–MS pair is assisted by one RS, but the same RS can serve more than one BS–MS pair. In fact, the number of BS–MSs pairs assisted by the rth RS equals the number of non zero entries of the rth row of the connectivity matrix $\mathbf{L}$.

We shall use the vectors $\boldsymbol{\ell}_1(t) \in \mathbb{R}_+^{\mathrm{M} \times 1}$, $\boldsymbol{\ell}_2(t) \in \mathbb{R}_+^{\mathrm{R} \times 1}$, and $\boldsymbol{\ell}_3(t) \in \mathbb{R}_+^{\mathrm{M} \times 1}$ to denote the CSI collected at the beginning of the tth frame. While $\ell_{1,m}(t)$ stands for the path loss between the BS and the mth MS, $\ell_{2,r}(t)$ is the path loss between the BS and the rth RS, and $\ell_{3,m}(t)$ is the path loss between the mth MS and its associated RS (which is the $\mathrm{RS}(m)$th). All of them are assumed to satisfy reciprocity.

When OFDM is employed with the only knowledge of the link path loss at each transmitter, one practical strategy is to perform uniform power allocation among groups of tones sufficiently far apart such that their individual fading states are uncorrelated and frequency diversity is enabled. This is the case in the IEEE 802.16e—PUSC and FUSC standards [14]. With this approach, coding across a sufficiently large number of tones makes the instantaneous achievable rate, denoted by $r(t)$, be upper bounded by the ergodic (or average) mutual information thanks to the law of large numbers.[3] By ergodic capacity we understand the instantaneous capacity given some fading state in the frequency domain averaged over all possible fading realizations in this domain. Therefore, no matter how short the transmission interval is nor how fast the channel response varies, the ergodic capacity will exclusively depend on the transmission bandwidth and the link signal-to-noise ratio,

---

[2]Although the proposed optimization framework can be extended to the FDD mode, we have ruled it out because it places more restrictive complexity requirements on the RSs, which should be able to receive and transmit simultaneously on different frequency bands.

[3]Consider a SISO point-to-point link for simplicity. When the transmit power is uniformly allocated over N tones spanning some total bandwidth B, the per-tone snr is constant. If $\{h_i\}_{i=1}^N$ denote the fading states of each tone (assumed i.i.d and unknown), the achievable rate satisfies

$$r(t) \leq \sum_{i=1}^N \frac{B}{N} \log_2(1 + \mathsf{snr} h_i)$$
$$= \frac{1}{N} \sum_{i=1}^N B \log_2(1 + \mathsf{snr} h_i) \xrightarrow{(a)} \mathbb{E}\{B \log_2(1 + \mathsf{snr} h)\},$$

where (a) follows for large N from the law of large numbers and convergence is in probability. For finite moderate values of N, outage events are not precluded. Its impact on system design, however, is beyond the scope of this paper.
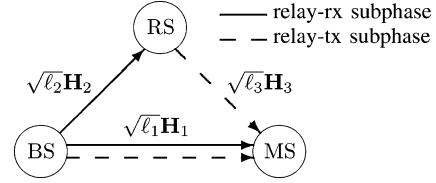

Fig. 1. DL cooperation protocol: the DL phase is split into two subphases attending to the half duplex nature of the RS.

snr. The snr suffices to characterize the quality of a link since interference between neighboring transmitters is prevented by allocating bandwidth among the different BS-RS-MS triplets in a *disjoint* manner[4]: each BS–MS–RS triplet is assigned a fraction of the total bandwidth in exclusivity. This fraction may vary from UL to DL phases and also within each of them, depending on whether the RS is active (relay-transmit subphase) or not (relay-receive subphase). Whichever subphase we focus on, the snr of any link is given by

$$\mathsf{snr} = \frac{\ell G p}{N_0 F b} \tag{1}$$

where ℓ is pathloss, G is antenna gain, p is transmit power, N_0 is the AWGN one-sided power spectral density, F is the noise factor, and b is bandwidth. Whereas the specific values of p and b are subject to optimization by the BS and N_0 and ℓ are given, we distinguish between $F_{\mathrm{BS}}(G_{\mathrm{BS}})$, $F_{\mathrm{RS}}(G_{\mathrm{RS}})$, and $F_{\mathrm{MS},m}(G_{\mathrm{MS},m})$ to consider the general case of nodes equipped with RF front-ends of diverse quality.

III. RELAY-ASSISTED TRANSMISSION

A. Maximum Instantaneous Achievable Rates

The use of RSs in our network setup has the advantage of realizing performance gains arising from relay-assisted transmission. As the bandwidth is assigned orthogonally (disjointly) to each BS-RS-MS triplet, intra-cell interference is completely nulled and it suffices to study one single triplet to describe the overall behavior of the cell.

Considering that every RS operates in the half duplex mode, then for a given time duration the relay is in the receive mode (we call this period the relay-receive subphase), and in the transmit mode for the rest (we call this period the relay-transmit subphase).[5] To illustrate the cooperation protocol, which is that of [6, Prop. 2], consider the specific BS-RS-MS triplet shown in Fig. 1, where the DL phase is described and the MS and RS index are omitted for simplicity. The matrices $\mathbf{H}_1 \in \mathbb{C}^{n_{\mathrm{MS}} \times n_{\mathrm{BS}}}$, $\mathbf{H}_2 \in \mathbb{C}^{n_{\mathrm{RS}} \times n_{\mathrm{BS}}}$, and $\mathbf{H}_3 \in \mathbb{C}^{n_{\mathrm{MS}} \times n_{\mathrm{RS}}}$ represent the instantaneous fading states of each of the links at a given tone.[6]

Although the details of the coding scheme can be found in [6, App. A], we provide here a brief sketch of it for the sake of clarity. The BS splits its message into two independent components: one which is transmitted directly to the MS without the

---

[4]Inter-cell interference due to frequency reuse in neighboring cells is not considered in this paper due to our single-cell approach.

[5]We shall also refer to the protocol subphases as subphase 1 (relay-receive) and subphase 2 (relay-transmit).

[6]When UL cooperative transmission is considered, the instantaneous fading states can be described by using the transposed matrices $\{\mathbf{H}_j^T\}_{j=1}^3$.

help of the RS, and another which is transmitted through the RS to the MS. During the relay-receive subphase, of duration τ_1, the BS transmits one codeword related to the latter message component using some power p_1 while the RS and the MS listen. At the end of this subphase, the RS attempts to decode this message component. If successful, a relay-transmit subphase of duration τ_2 starts where both the RS (which re-encodes the decoded message component) and the BS (which now transmits a codeword associated to its other message component) transmit using powers p_3 and p_2, respectively. Otherwise the RS remains silent during this subphase. The receiver performs successive decoding: it first attempts to decode the relayed message component from the signal of both subphases and, if successful, it subtracts the signal transmitted by the RS in the second subphase prior to decoding the unrelayed message component. Assuming that communication takes place over bandwidths B_1 (relay-receive subphase) and B_2 (relay-transmit subphase) and that uniform power allocation across antennas is performed (see [28] for relay-assisted communication protocols where power allocation is performed assuming perfect CSI), the achievable rate r_{DL} in [bit/s] satisfies [6]

$$r_{\mathrm{DL}} \leq \min\left\{r_{\mathrm{DL}}^{(1)}, r_{\mathrm{DL}}^{(2)}\right\} \qquad (2)$$

where the min function models whether it is the source-relay link or the source-destination link which act as information bottlenecks for the relayed message component, and

$$r_{\mathrm{DL}}^{(1)} = \tau_1 B_1 \mathbb{E}\left\{\log_2 \det\left(\mathbf{I}_{n_{\mathrm{RS}}} + \frac{\mathsf{snr}_2}{n_{\mathrm{BS}}}\mathbf{H}_2\mathbf{H}_2^\dagger\right)\right\}$$
$$+ \tau_2 B_2 \mathbb{E}\left\{\log_2 \det\left(\mathbf{I}_{n_{\mathrm{MS}}} + \frac{\mathsf{snr}_{1,2}}{n_{\mathrm{BS}}}\mathbf{H}_1\mathbf{H}_1^\dagger\right)\right\} \quad (3)$$

$$r_{\mathrm{DL}}^{(2)} = \tau_1 B_1 \mathbb{E}\left\{\log_2 \det\left(\mathbf{I}_{n_{\mathrm{MS}}} + \frac{\mathsf{snr}_{1,1}}{n_{\mathrm{BS}}}\mathbf{H}_1\mathbf{H}_1^\dagger\right)\right\}$$
$$+ \tau_2 B_2 \mathbb{E}\left\{\log_2 \det\left(\mathbf{I}_{n_{\mathrm{MS}}} + \frac{\mathsf{snr}_{1,2}}{n_{\mathrm{BS}}}\mathbf{H}_1\mathbf{H}_1^\dagger\right.\right.$$
$$\left.\left. + \frac{\mathsf{snr}_3}{n_{\mathrm{RS}}}\mathbf{H}_3\mathbf{H}_3^\dagger\right)\right\} \quad (4)$$

where the snrs amount to

$$\mathsf{snr}_{1,j} = \frac{\ell_1 G_{\mathrm{BS}} p_j}{N_0 F_{\mathrm{MS}} B_j}, \quad \mathsf{snr}_2 = \frac{\ell_2 G_{\mathrm{BS}} p_1}{N_0 F_{\mathrm{RS}} B_1}, \quad \mathsf{snr}_3 = \frac{\ell_3 G_{\mathrm{RS}} p_3}{N_0 F_{\mathrm{MS}} B_2} \qquad (5)$$

where $j = 1, 2$.

While the success of decoding the relayed message component at the relay indeed impacts on the success of decoding at the destination, the behavior of the destination is independent of whether the relay was able to decode or not. The destination will attempt to decode first the relayed component, perform successive interference cancellation, and go for the direct component afterwards, no matter what happened to the relay. This makes the relay a transparent network feature as seen by the MS, as no signalling between them is required whatsoever. In fact, as we rely on the ergodic capacities to characterize performance (see Section II), it can be assumed that all the transmissions are reliable as long as their information rates lie below capacity. Consequently, the performance (2) of the strategy [6] for the one-way

relay channel with half-duplex relay is such that the transmission rate of the relayed message component always results in successful decoding at the relay.

It is important to remark that the upper bound (2) is only tight for Gaussian codes of infinite blocklength. When practical discrete alphabet codes of finite blocklength are used instead, decoding errors at the RS and the MS cannot be disregarded at rates below the corresponding ergodic capacities. However, expression (2) can still be used by introducing a penalizing gap Γ such that $\mathsf{snr}_{\mathrm{practical}} = \mathsf{snr}/\Gamma$ in (3) and (4).[7] We will hence use the gap from now on and omit the subscript "practical" in snr for simplicity. When UL transmission is considered, an analogous expression to (2) of the form $r_{\mathrm{UL}} \leq \min\{r_{\mathrm{UL}}^{(1)}, r_{\mathrm{UL}}^{(2)}\}$ readily follows by exchanging the roles of the MS and the BS and transposing the matrices $\{\mathbf{H}_j\}_{j=1}^3$.

Oppositely to [21], where relays explicitly switched between amplify-and-forward and decode-and-forward depending on the achievable rates, the adopted cooperation protocol has the advantage of comprising other well-known cooperation strategies as particular cases such that the best one is implicitly selected when the resource allocation is optimized. While it mimics the philosophy of protocol I of [3] and transmit diversity [5], it can also accommodate the following:

- *Protocol III* [3], *simplified transmit diversity* [5]—Set p_1 to be too small to enable direct BS–MS reliable communication in the relay-receive subphase.
- *Protocol II* [3], *receive diversity* [3]—Set $p_2 = 0$.
- *Multihop relaying* [5], [29], [30]—Set $p_2 = 0$ and p_1 to be too small to enable direct BS–MS reliable communication in the relay-receive subphase.
- *Direct transmission*—Set $p_3 = 0$ and/or $\tau_2 = 0$ and/or $B_2 = 0$.

B. Universal Concave Lower Bounds on the Achievable Rates

Transmission over multiple tones with uncorrelated fading makes the ergodic (or average) rates show up in (3) and (4). They involve computing three MIMO channel ergodic capacities and one distributed MIMO channel ergodic capacity [the τ_2 term in (4)]. After averaging over the fading distribution, i.e., the distribution of the matrices $\{\mathbf{H}_j\}_{j=1}^3$, the resulting expectations depend only on the link snrs and the $\tau_j B_j$ products, and admit closed-form expressions for both the MIMO [10], [31] and the distributed MIMO [32] channel in case of Rayleigh fading. However, analytical expressions cannot be derived for other fading distributions like Ricean, that are common in the BS-RS link and include line-of-sight components (LOS). On top of that, (3) and (4) are not concave functions of the duration of the subphases, the allocated bandwidths, and the transmit powers. This prevents efficient methods to be applied for rate allocation in global optimization approaches.

Alternatively, we develop universal, simpler concave lower bounds of $r_{\mathrm{DL}}^{(1)}$ and $r_{\mathrm{DL}}^{(2)}$ that ease prospective optimization methods and allow for an easy concavity test. Here, by "universal," we mean that parametric lower bounds with the same *structure* can be applied to *any* fading distribution by changing

---

[7]The gap can be further increased to model the impact of inter-cell interference on final performance via snr degradation.

the parameter values and not that the same expression holds for all of them. Other parametric approaches have been taken to approximate MIMO ergodic capacities [33], but oppositely to our needs, concavity with respect to durations, bandwidths, and powers was not guaranteed, parameter values were not systematically found (i.e., curve fitting was performed), and the distributed MIMO case was not tackled. To start with, consider the following results upon which our lower bounds are based. Their concavity analysis will be left to the next section.

Lemma 1: A lower bound to the ergodic capacity of an $n_t \times n_r$ MIMO channel is

$$\mathbb{E}\left\{\log_2 \det\left(\mathbf{I}_{n_r} + \frac{\mathsf{snr}}{n_t}\mathbf{H}\mathbf{H}^\dagger\right)\right\} \geq \sum_{i=1}^{n_r} \log_2(1 + \rho_i(f_{\mathbf{H}})\mathsf{snr}/n_t) \tag{6}$$

where $\rho_i(f_{\mathbf{H}}) \triangleq \exp(\mathbb{E}\{\log \lambda_i(\mathbf{H}\mathbf{H}^\dagger))\})$ and $f_{\mathbf{H}}(\cdot)$ denotes the pdf of the channel matrix $\mathbf{H}$.[8]

Proof: Proceeding as in [10, App. E.1], we start from the expression that relates the ergodic capacity to the ordered eigenvalues of $\mathbf{H}\mathbf{H}^\dagger$ to obtain

$$\mathbb{E}\left\{\log_2 \det\left(\mathbf{I}_{n_r} + \frac{\mathsf{snr}}{n_t}\mathbf{H}\mathbf{H}^\dagger\right)\right\}$$

$$= \sum_{i=1}^{n_r} \mathbb{E}\left\{\log_2\left(1 + \lambda_i(\mathbf{H}\mathbf{H}^\dagger)\mathsf{snr}/n_t\right)\right\} \tag{7}$$

$$= \sum_{i=1}^{n_r} \mathbb{E}\left\{\log_2\left(1 + \exp\left(\log \lambda_i(\mathbf{H}\mathbf{H}^\dagger)\right)\mathsf{snr}/n_t\right)\right\} \tag{8}$$

$$\geq \sum_{i=1}^{n_r} \log_2\left(1 + \exp\left(\mathbb{E}\left\{\log \lambda_i(\mathbf{H}\mathbf{H}^\dagger)\right\}\right)\mathsf{snr}/n_t\right) \tag{9}$$

where (9) follows from Jensen's inequality and the convexity of the function $I(x) = \log_2(a + b\exp(x))$ for all $a, b \geq 0$. ∎

Lemma 2: A lower bound to the ergodic capacity of an $n_{t,1} \times n_r$ and $n_{t,2} \times n_r$ distributed MIMO channel is

$$\mathbb{E}\left\{\log_2 \det\left(\mathbf{I}_{n_r} + \frac{\mathsf{snr}_1}{n_{t,1}}\mathbf{H}_1\mathbf{H}_1^\dagger + \frac{\mathsf{snr}_2}{n_{t,2}}\mathbf{H}_2\mathbf{H}_2^\dagger\right)\right\}$$

$$\geq \sum_{i=1}^{n_r} \log_2(1 + \rho_i(f_{\mathbf{H}_1})\mathsf{snr}_1/n_{t,1} + \rho_i(f_{\mathbf{H}_2})\mathsf{snr}_2/n_{t,2}). \tag{10}$$

Proof: Since $1 + \lambda_i(\mathbf{H}_1\mathbf{H}_1^\dagger)\mathsf{snr}_1/n_{t,1} + \lambda_i(\mathbf{H}_2\mathbf{H}_2^\dagger)\mathsf{snr}_2/n_{t,2} \geq 0$ for $1 \leq i \leq n_r$ and both $\mathbf{H}_1\mathbf{H}_1^\dagger$ and $\mathbf{H}_2\mathbf{H}_2^\dagger$ are Hermitian matrices, it follows from [34] that

$$\mathbb{E}\left\{\log_2 \det\left(\mathbf{I}_{n_r} + \frac{\mathsf{snr}_1}{n_{t,1}}\mathbf{H}_1\mathbf{H}_1^\dagger + \frac{\mathsf{snr}_2}{n_{t,2}}\mathbf{H}_2\mathbf{H}_2^\dagger\right)\right\}$$

$$\geq \sum_{i=1}^{n_r} \mathbb{E}\left\{\log_2\left(1 + \lambda_i\left(\mathbf{H}_1\mathbf{H}_1^\dagger\right)\mathsf{snr}_1/n_{t,1}\right.\right.$$

$$\left.\left. + \lambda_i\left(\mathbf{H}_2\mathbf{H}_2^\dagger\right)\mathsf{snr}_2/n_{t,2}\right)\right\}. \tag{11}$$

The lemma follows by similarly applying Jensen's inequality resorting twice to the function $I(x)$. ∎

[8]Note that since $\text{rank}\{\mathbf{H}\mathbf{H}^\dagger\} \leq \min\{n_t, n_r\}$, $\rho_i(f_{\mathbf{H}}) = 0$ for $\min\{n_t, n_r\} < i \leq n_r$.
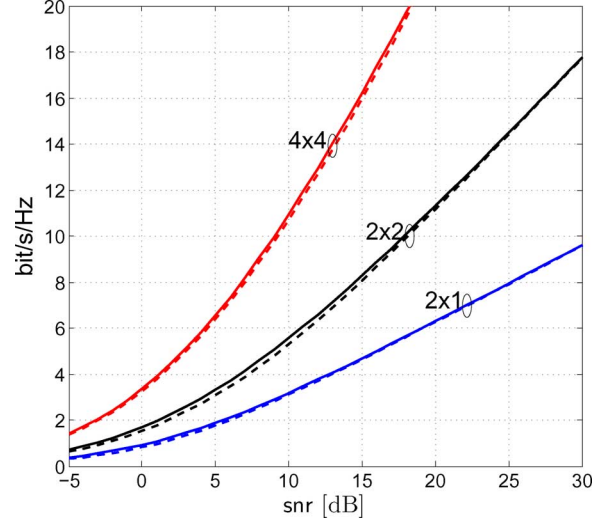

Fig. 2. Exact ergodic capacity (solid lines) and Lemma 1 lower bound (dashed lines) versus snr for different antenna configurations and Rayleigh fading.
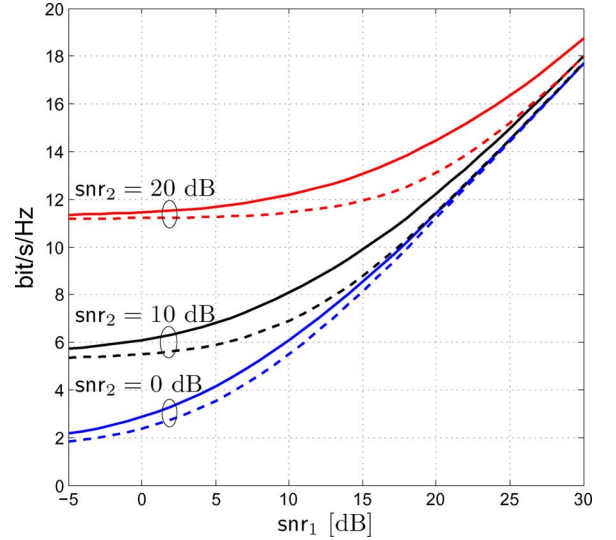
Fig. 3. Exact ergodic capacity (solid lines) and Lemma 2 lower bound (dashed lines) versus snr_1 for different values of snr_2 and Rayleigh fading. The antenna configuration is $n_r = n_{t,1} = n_{t,2} = 2$.

Lemmas 1 and 2 lower bound the MIMO channel capacities with expressions that mimic equivalent transmissions through virtual parallel AWGN channels of gains $\rho_i(\cdot)$, which depend on the antenna configuration and the fading distribution, and whose tightness is analyzed in Figs. 2 and 3. As for the MIMO channel, Lemma 1 lower bound is extremely tight. The tightness of Lemma 2 lower bound with respect to the distributed MIMO channel capacity, however, depends on the snr.

The computation of the channel-dependent coefficients $\rho_i(\cdot)$ can be accurately performed offline by using Monte Carlo methods. However, as for Rayleigh fading and an $n \times 1$ or $1 \times n$ antenna configuration, results on the expectation of the logarithm of a Chi-square random variable [35] can be applied to show that

$$\rho_i(n) = e^{-\Psi + \sum_{j=1}^{n-1} \frac{1}{j}}\delta[i-1] \tag{12}$$

where $\Psi \approx 0.577$ is the Euler–Mascheroni constant [36, 4.352-1]. In any case, once the channel-dependent coefficients are computed, Lemmas 1 and 2 allow us to state the main result of this section.

Corollary 1: A lower bound on the maximum DL achievable rates of the adopted relay-assisted transmission protocol is

$$r_{\mathrm{DL}} \leq \min\left\{\bar{r}_{\mathrm{DL}}^{(1)}, \bar{r}_{\mathrm{DL}}^{(2)}\right\} \tag{13}$$

where

$$\bar{r}_{\mathrm{DL}}^{(1)} = \tau_1 B_1 \sum_{i=1}^{n_{\mathrm{RS}}} \log_2\left(1 + \rho_i\left(f_{\mathbf{H}_2}\right)\frac{\mathsf{snr}_2}{n_{\mathrm{BS}}}\right)$$
$$+ \tau_2 B_2 \sum_{i=1}^{n_{\mathrm{MS}}} \log_2\left(1 + \rho_i\left(f_{\mathbf{H}_1}\right)\frac{\mathsf{snr}_{1,2}}{n_{\mathrm{BS}}}\right) \leq r_{\mathrm{DL}}^{(1)} \tag{14}$$

$$\bar{r}_{\mathrm{DL}}^{(2)} = \tau_1 B_1 \sum_{i=1}^{n_{\mathrm{MS}}} \log_2\left(1 + \rho_i\left(f_{\mathbf{H}_1}\right)\frac{\mathsf{snr}_{1,1}}{n_{\mathrm{BS}}}\right)$$
$$+ \tau_2 B_2 \sum_{i=1}^{n_{\mathrm{MS}}} \log_2\left(1 + \rho_i\left(f_{\mathbf{H}_1}\right)\frac{\mathsf{snr}_{1,2}}{n_{\mathrm{BS}}}\right.$$
$$\left. + \rho_i\left(f_{\mathbf{H}_3}\right)\frac{\mathsf{snr}_3}{n_{\mathrm{RS}}}\right) \leq r_{\mathrm{DL}}^{(2)}. \tag{15}$$

A similar lower bound on the maximum UL achievable rates holds by exchanging the roles of the BS and the MS and transposing $\{\mathbf{H}_j\}_{j=1}^3$.

IV. ACHIEVABLE INSTANTANEOUS RATES

Given the CSI at the beginning of the tth frame, $\{\boldsymbol{\ell}_j(t)\}_{j=1}^3$, the instantaneous performance of the network is given by the DL and UL achievable rate regions, i.e., the set of all rate vectors $\mathbf{r}_{\mathrm{DL}}(t), \mathbf{r}_{\mathrm{UL}}(t) \in \mathbb{R}_+^{\mathsf{M}\times 1}$ that can be sustained during one frame duration. The achievable rates depend upon the frame format as described by the vector of fractional durations $\boldsymbol{\tau} \in \mathbb{R}_+^4$, $\mathbf{1}_4^T \boldsymbol{\tau} = 1$, whose components account for the DL subphase 1 (τ_1), DL subphase 2 (τ_2), UL subphase 1 (τ_3), and UL subphase 2 (τ_4). Each subphase duration shapes and couples the instantaneous achievable rate regions, denoted by $\mathcal{R}_{\mathrm{DL}}(t; \boldsymbol{\tau})$ (DL) and $\mathcal{R}_{\mathrm{UL}}(t; \boldsymbol{\tau})$ (UL), and will be subject to optimization later on, when rate allocation policies come into play in the next section. In this section, however, we shall focus on the dependence of the achievable rates on the disjoint allocations of power among transmitters and bandwidth among BS-RS-MS triplets.

A. DL Instantaneous Achievable Rate Region

Assuming that the duration of the DL subphases is fixed to $\tau_1 T$ and $\tau_2 T$, the instantaneous achievable rates depend upon the allocation of bandwidth and transmit power among the M competing flows. Let us describe the DL resource allocation by using the vectors $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}_+^{\mathsf{M}\times 1}$, which represent the fractional BS power allocation in subphases 1 and 2, the fractional RS's power allocation in subphase 2 ($p_{3,m}$ is the fraction of power transmitted by the $\mathrm{RS}(m)$th RS in assisting the mth MS), and the fractional bandwidth allocation in subphases 1 and

2, respectively. By imposing non-negativity on each fraction and constraining the sum of resources it follows that

$$\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{b}_1, \mathbf{b}_2 \geq \mathbf{0}_{\mathsf{M}} \tag{16}$$
$$\mathbf{1}_{\mathsf{M}}^T \mathbf{p}_j \leq 1, \quad \mathbf{L}\mathbf{p}_3 \leq \mathbf{1}_{\mathsf{R}}, \quad \mathbf{1}_{\mathsf{M}}^T \mathbf{b}_j \leq 1 \tag{17}$$

where $j = 1, 2$ and $\mathbf{L}$ is the connectivity matrix defined in Section II. Thus, applying Corollary 1 the DL achievable rate in [bit/s] of the mth user, $r_{\mathrm{DL},m}(t)$, satisfies

$$r_{\mathrm{DL},m}(t) \leq \min\left\{\bar{r}_{\mathrm{DL},m}^{(1)}(t), \bar{r}_{\mathrm{DL},m}^{(2)}(t)\right\} \tag{18}$$

where

$$\bar{r}_{\mathrm{DL},m}^{(1)}(t)$$
$$= B\left[\tau_1 b_{1,m} \sum_{i=1}^{n_{\mathrm{RS}}} \log_2\left(1 + c_{2,m}^i(t)\frac{p_{1,m}}{b_{1,m}}\right)\right.$$
$$\left. + \tau_2 b_{2,m} \sum_{i=1}^{n_{\mathrm{MS},m}} \log_2\left(1 + c_{1,m}^i(t)\frac{p_{2,m}}{b_{2,m}}\right)\right] \tag{19}$$

$$\bar{r}_{\mathrm{DL},m}^{(2)}(t)$$
$$= B\left[\tau_1 b_{1,m} \sum_{i=1}^{n_{\mathrm{MS},m}} \log_2\left(1 + c_{1,m}^i(t)\frac{p_{1,m}}{b_{1,m}}\right)\right.$$
$$\left. + \tau_2 b_{2,m} \sum_{i=1}^{n_{\mathrm{MS},m}} \log_2\left(1 + \frac{c_{1,m}^i(t)p_{2,m} + c_{3,m}^i(t)p_{3,m}}{b_{2,m}}\right)\right] \tag{20}$$

condense CSI into the equivalent channel gains

$$c_{1,m}^i(t) = \frac{\rho_i\left(f_{\mathbf{H}_{1,m}}\right)\ell_{1,m}(t)G_{\mathrm{BS}}p_{\mathrm{BS}}^{\max}}{n_{\mathrm{BS}}\Gamma N_0 F_{\mathrm{MS}}B} \tag{21}$$

$$c_{2,m}^i(t) = \frac{\rho_i\left(f_{\mathbf{H}_{2,\mathrm{RS}(m)}}\right)\ell_{2,\mathrm{RS}(m)}(t)G_{\mathrm{BS}}p_{\mathrm{BS}}^{\max}}{n_{\mathrm{BS}}\Gamma N_0 F_{\mathrm{RS}}B} \tag{22}$$

$$c_{3,m}^i(t) = \frac{\rho_i\left(f_{\mathbf{H}_{3,m}}\right)\ell_{3,m}(t)G_{\mathrm{RS}}p_{\mathrm{RS}}^{\max}}{n_{\mathrm{RS}}\Gamma N_0 F_{\mathrm{MS}}B}. \tag{23}$$

Following the notation of Section III, we have used $f_{\mathbf{H}_{1,m}}$ to denote the DL fading distribution between the BS and the mth MS, $f_{\mathbf{H}_{2,\mathrm{RS}(m)}}$ for the DL fading distribution between the BS and the serving RS of the mth MS, and $f_{\mathbf{H}_{3,m}}$ for the DL fading distribution between the mth MS and its serving RS. The DL achievable rate region is hence given by

$$\mathcal{R}_{\mathrm{DL}}(t; \boldsymbol{\tau}) = \bigcup\left\{\mathbf{0}_{\mathsf{M}} \leq \mathbf{r}_{\mathrm{DL}}(t) \leq \min\left\{\bar{\mathbf{r}}_{\mathrm{DL}}^{(1)}(t), \bar{\mathbf{r}}_{\mathrm{DL}}^{(2)}(t)\right\}\right\} \tag{24}$$

where the union is taken over the allocations satisfying (16) and (17).

Lemma 3: The DL instantaneous achievable rate region $\mathcal{R}_{\mathrm{DL}}(t; \boldsymbol{\tau})$ is convex.

Proof: For fixed τ_1, τ_2, some properties of convex functions [37] can be used to show that the right-hand side of (18) is concave: the minimum of concave functions is concave, and the concavity of (19) and (20) can be shown resorting to the function $G(x, y) = ax\log(1 + by/x)$, which is concave in $x, y \geq 0$ $\forall a, b \geq 0$. This implies convexity of $\mathcal{R}_{\mathrm{DL}}(t; \boldsymbol{\tau})$ for fixed $\boldsymbol{\tau}$ [37], which will turn out to be useful in ensuring global

optimality in rate allocation problems. This desirable property, which follows from the use of the universal lower bounds derived in Section III-B, vanishes when the frame format (here in the form of the relative durations τ_1, τ_2) is subject to optimization too. However, this can be circumvented with the following variable change:

$$\mathbf{q}_j \triangleq \tau_j \mathbf{p}_j, \quad \mathbf{q}_3 \triangleq \tau_2 \mathbf{p}_3, \quad \mathbf{w}_j \triangleq \tau_j \mathbf{b}_j \qquad (25)$$

where $j = 1, 2$. This variable change gives rise to a new set of allocation variables, in terms of which (19), (20) become

$$
\begin{aligned}
&\overline{r}_{\mathrm{DL},m}^{(1)}(t) \\
&= B\left[w_{1,m} \sum_{i=1}^{n_{\mathrm{RS}}} \log_2\left(1 + c_{2,m}^i(t)\frac{q_{1,m}}{w_{1,m}}\right) \right. \\
&\left. \quad + w_{2,m} \sum_{i=1}^{n_{\mathrm{MS},m}} \log_2\left(1 + c_{1,m}^i(t)\frac{q_{2,m}}{w_{2,m}}\right) \right]
\end{aligned}
\qquad (26)
$$

$$
\begin{aligned}
&\overline{r}_{\mathrm{DL},m}^{(2)}(t) \\
&= B\left[w_{1,m} \sum_{i=1}^{n_{\mathrm{MS},m}} \log_2\left(1 + c_{1,m}^i(t)\frac{q_{1,m}}{w_{1,m}}\right) \right. \\
&\left. \quad + w_{2,m} \sum_{i=1}^{n_{\mathrm{MS},m}} \log_2\left(1 + \frac{c_{1,m}^i(t)q_{2,m} + c_{3,m}^i(t)q_{3,m}}{w_{2,m}}\right) \right]
\end{aligned}
\qquad (27)
$$

both of them concave functions regardless of τ_1, τ_2. The new set of feasible allocations transforms accordingly into

$$\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \mathbf{w}_1, \mathbf{w}_2 \geq \mathbf{0}_{\mathrm{M}} \qquad (28)$$

$$\mathbf{1}_{\mathrm{M}}^T \mathbf{q}_j \leq \tau_j, \quad \mathbf{L}\mathbf{q}_3 \leq \tau_2 \mathbf{1}_{\mathrm{R}}, \quad \mathbf{1}_{\mathrm{M}}^T \mathbf{w}_j \leq \tau_j \qquad (29)$$

where $j = 1, 2$ again. Formulated in terms of the new allocation variables, the region $\mathcal{R}_{\mathrm{DL}}(t; \boldsymbol{\tau})$ can be equivalently obtained by taking the union in (24) over (28) and (29), where (26) and (27) are used instead of (19) and (20). This way, the convexity of $\mathcal{R}_{\mathrm{DL}}(t; \boldsymbol{\tau})$ with respect to $\boldsymbol{\tau}$ is unveiled: (26) and (27) are concave and the feasible set (28) and (29) is the intersection of halfspaces and hence convex, something that was hidden with the original allocation variables. Needless to say, the variable change (25) can be straightforwardly reversed to obtain the allocated fractions of bandwidth and power. ∎

B. UL Instantaneous Achievable Rate Region

Proceeding similarly, if the relative duration of the subphases is fixed to $\tau_3 T$ and $\tau_4 T$, the UL resource allocation can be characterized in terms of the vectors $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}_+^{\mathrm{M}}$. While the meaning of $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{p}_3$ is identical, $\mathbf{p}_1$ and $\mathbf{p}_2$ refer now to the fractional MS's transmit power in subphases 1 and 2. Thus, the feasible set of UL resource allocations is

$$\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{b}_1, \mathbf{b}_2 \geq \mathbf{0}_{\mathrm{M}} \qquad (30)$$

$$\mathbf{p}_j \leq \mathbf{1}_{\mathrm{M}}, \quad \mathbf{L}\mathbf{p}_3 \leq \mathbf{1}_{\mathrm{R}}, \quad \mathbf{1}_{\mathrm{M}}^T \mathbf{b}_j \leq 1 \qquad (31)$$

where $j = 1, 2$. The application of Corollary 1 to the UL achievable rate in [bit/s] of the mth user implies that $r_{\mathrm{UL},m}(t)$ satisfies

$$r_{\mathrm{UL},m}(t) \leq \min\left\{ \overline{r}_{\mathrm{UL},m}^{(1)}(t), \overline{r}_{\mathrm{UL},m}^{(2)}(t) \right\} \qquad (32)$$

where

$$
\begin{aligned}
&\overline{r}_{\mathrm{UL},m}^{(1)}(t) \\
&= B\left[\tau_1 b_{1,m} \sum_{i=1}^{n_{\mathrm{RS}}} \log_2\left(1 + d_{3,m}^i(t)\frac{p_{1,m}}{b_{1,m}}\right) \right. \\
&\left. \quad + \tau_2 b_{2,m} \sum_{i=1}^{n_{\mathrm{BS},m}} \log_2\left(1 + d_{1,m}^i(t)\frac{p_{2,m}}{b_{2,m}}\right) \right]
\end{aligned}
\qquad (33)
$$

$$
\begin{aligned}
&\overline{r}_{\mathrm{UL},m}^{(2)}(t) \\
&= B\left[\tau_1 b_{1,m} \sum_{i=1}^{n_{\mathrm{BS},m}} \log_2\left(1 + d_{1,m}^i(t)\frac{p_{1,m}}{b_{1,m}}\right) \right. \\
&\left. \quad + \tau_2 b_{2,m} \sum_{i=1}^{n_{\mathrm{BS},m}} \log_2\left(1 + \frac{d_{1,m}^i(t)p_{2,m} + d_{2,m}^i(t)p_{3,m}}{b_{2,m}}\right) \right]
\end{aligned}
\qquad (34)
$$

use the equivalent channel gains

$$d_{1,m}^i(t) = \frac{\rho_i\left(f_{\mathbf{H}_{1,m}^T}\right) \ell_{1,m}(t) G_{\mathrm{MS}} p_{\mathrm{MS}}^{\max}}{n_{\mathrm{MS}} \Gamma N_0 F_{\mathrm{BS}} B} \qquad (35)$$

$$d_{2,m}^i(t) = \frac{\rho_i\left(f_{\mathbf{H}_{2,\mathrm{RS}(m)}^T}\right) \ell_{2,\mathrm{RS}(m)}(t) G_{\mathrm{RS}} p_{\mathrm{RS}}^{\max}}{n_{\mathrm{RS}} \Gamma N_0 F_{\mathrm{BS}} B} \qquad (36)$$

$$d_{3,m}^i(t) = \frac{\rho_i\left(f_{\mathbf{H}_{3,m}^T}\right) \ell_{3,m}(t) G_{\mathrm{MS}} p_{\mathrm{MS}}^{\max}}{n_{\mathrm{MS}} \Gamma N_0 F_{\mathrm{RS}} B}. \qquad (37)$$

Note that by transposing the DL fading state matrices, the fading distributions account for UL transmission. The UL achievable rate region can finally be expressed as

$$\mathcal{R}_{\mathrm{UL}}(t; \boldsymbol{\tau}) = \bigcup\left\{ \mathbf{0}_{\mathrm{M}} \leq \mathbf{r}_{\mathrm{UL}}(t) \leq \min\left\{ \overline{\mathbf{r}}_{\mathrm{UL}}^{(1)}(t), \overline{\mathbf{r}}_{\mathrm{UL}}^{(2)}(t) \right\} \right\} \qquad (38)$$

where the union is over (30) and (31).

Lemma 4: The UL instantaneous achievable rate region $\mathcal{R}_{\mathrm{UL}}(t; \boldsymbol{\tau})$ is convex.

Proof: As happened in the DL, the achievable rate region $\mathcal{R}_{\mathrm{UL}}(t; \boldsymbol{\tau})$ is convex when the frame format $\boldsymbol{\tau}$ is fixed, but not when it becomes an optimization variable. To avoid this handicap, the same variable change as in (25) is proposed. This leads to the concave expressions

$$
\begin{aligned}
&\overline{r}_{\mathrm{UL},m}^{(1)}(t) \\
&= B\left[w_{1,m} \sum_{i=1}^{n_{\mathrm{RS}}} \log_2\left(1 + d_{3,m}^i(t)\frac{q_{1,m}}{w_{1,m}}\right) \right. \\
&\left. \quad + w_{2,m} \sum_{i=1}^{n_{\mathrm{BS},m}} \log_2\left(1 + d_{1,m}^i(t)\frac{q_{2,m}}{w_{2,m}}\right) \right]
\end{aligned}
\qquad (39)
$$

$$
\begin{aligned}
&\overline{r}_{\mathrm{UL},m}^{(2)}(t) \\
&= B\left[w_{1,m} \sum_{i=1}^{n_{\mathrm{BS},m}} \log_2\left(1 + d_{1,m}^i(t)\frac{q_{1,m}}{w_{1,m}}\right) \right. \\
&\left. \quad + w_{2,m} \sum_{i=1}^{n_{\mathrm{BS},m}} \log_2\left(1 + \frac{d_{1,m}^i(t)q_{2,m} + d_{2,m}^i(t)q_{3,m}}{w_{2,m}}\right) \right]
\end{aligned}
\qquad (40)
$$

and the new feasible set

$$\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \mathbf{w}_1, \mathbf{w}_2 \geq \mathbf{0}_M \tag{41}$$

$$\mathbf{q}_j \leq \tau_j \mathbf{1}_M, \quad \mathbf{L}\mathbf{q}_3 \leq \tau_2 \mathbf{1}_R, \quad \mathbf{1}_M^T \mathbf{w}_j \leq 1 \tag{42}$$

where $j = 1, 2$. Thus, a convex representation of $\mathcal{R}_{\mathrm{UL}}(t; \boldsymbol{\tau})$ follows by replacing (33) and (34) by (39) and (40) in (38) and performing the union over the allocations satisfying (41) and (42). ∎

V. MAXIMUM NETWORK UTILITY RATE ALLOCATION POLICIES

In a cellular network handling users of services with different QoS requirements, some users might exhibit large sensitivity to transmission delays while others might only pay attention to their experienced long-term throughput. And still other key performance indicators may play a role, such as transmit buffer overflow probability, energy consumption… etc; even users of the same service, attending to their contract terms, can ask for differentiated QoS requirements.

This poses a challenging problem: upon collection of CSI at the beginning of the tth frame, the BS must decide by whom, when, and at which rate any information will be transmitted/received until the next CSI refresh arrives. Assuming that CSI updates are received periodically every D frames, user scheduling and resource allocation needs to be jointly optimized for UL and DL transmission during the frames $\{t, t+1, \ldots, t+D-1\}$. On the one hand, flows of different nature may require completely different management policies; on the other, the network operation should seamlessly reconfigure as the scenario (users, services, QoS requirements, etc.) varies with time.

We address this problem by using utility functions [21], [24] that evaluate each user's satisfaction given the achieved throughput as compared to its requirements. By properly characterizing QoS requirements with the dependency of the utility on the throughput for each service under operation, the different nature of the serving flows is incorporated into the BS arbitrage. An arbitrage that will use the utility function of each user to allocate resources and perform scheduling decisions.

A. User Utility Functions

Utility functions were first used in [24] to introduce the proportional fair criterion in resource allocation problems, and allow us to describe the satisfaction of one user given its served throughput. Although many schemes implicitly assume utilities proportional to throughputs [17], [19], [30], we shall adopt here a more general approach as in [20], [21], and [24]–[27]. We define a user utility function $U(R)$ as a concave function of the (long-term) throughput R, which is computed using the exponentially weighted smoothing

$$R(t+1) = \lambda R(t) + (1-\lambda) r(t) \tag{43}$$

where $R(t+1)$ stands for the throughput as seen at the beginning of the $(t+1)$-frame, λ represents the smoothing memory, and $r(t)$ is the instantaneous rate achieved in the tth frame. If for any reason the user satisfaction profile of a service cannot be described using a concave function (e.g., an S-type curve),

the finding of efficient methods for network utility maximization is compromised. Fortunately, a plethora of common services comply with the concavity constraint.

1) QoS-Oriented Utility Functions: We say a utility function is QoS-oriented whenever the QoS requirements of the service appear explicitly in its expression. Although we are not constrained to it for operational reasons, we focus on utility functions upper-bounded by 1. This way, we set the same maximum user satisfaction level as a reference for all the services under operation in the network. Otherwise, the use of unbounded utilities for the different services might cause the BS to bias its attention towards services with favored utility scales. The following examples show that this is not a major impairment to describe the satisfaction profile of services of different nature.

- *Example 1: Best-effort data service*
  The user satisfaction of a best-effort data service (e.g., ftp, http) without any data latency or other QoS constraints than achieving the largest possible throughput can be modeled with the utility function

$$U(R) = 1 - e^{\log(1-U_0)\frac{R}{R_0}} = 1 - (1 - U_0)^{\frac{R}{R_0}}. \tag{44}$$

  This utility is parameterized by the satisfaction level $0 < U_0 < 1$ achieved when the throughput is R_0.

- *Example 2: Delay-sensitive service*
  The user of a delay-sensitive service is interested in achieving some target throughput under the constraint that data latency remains below some critical threshold: in practical applications (e.g., voice service, video streaming), bits exceeding the maximum allowed delay are dropped. Since such applications are usually of constant bit rate, allocation of rates larger than the target throughput renders suboptimal. In other words, an overuse of resources makes no real improvement for this user but compromises the QoS provision to the rest. This can be alternatively viewed in terms of imposing the instantaneous rate to be as constant as possible, thus avoiding bursty transmissions yielding the same throughput at the expense of larger idle periods (and hence delays). One suitable utility function is

$$U(R) = 1 - \left(\frac{R - R_0}{\sigma}\right)^2 \tag{45}$$

  where R_0 is the target throughput and σ depends on the maximum allowable delay W_0 (in number of frames). To select σ, we choose to set the utility of one user that was initially served R_0 but is laid aside W_0 frames idle equal to $U_0 \ll 1$. This forces the satisfaction index of this user to move from the peak of (45), where utility was 1, to some unacceptable value U_0. This way, each frame one such user is idle it is able to warn the BS about its urgency for being scheduled by decreasing its utility. With this criterion, the appropriate σ is

$$\sigma = \frac{(1 - \lambda^{W_0}) R_0}{\sqrt{1 - U_0}}. \tag{46}$$

In case users of several delay-sensitive services with different QoS requirements (as specified by R_0 and W_0) are

present in the network, one has only to adjust σ according to (46) and use the resulting utility function (45).

2) Best-Effort Utility Functions: In situations where the utility function of a service does not depend on the QoS requirements, we say that utility is best-effort. As we are not able to quantify how far we are from the user expectations, we rather use utility as a qualitative satisfaction index. In addition, if there is only one service under operation in the network, there is no reason to focus on functions upper-bounded by 1. This can allow us to consider a wider class of utility functions. A useful example of best-effort utility function is the family [38]

$$U(R) = \begin{cases} \log(R) & \text{if } \alpha = 1 \\ \frac{R^{1-\alpha}}{1-\alpha} & \text{if } \alpha \neq 1 \end{cases} \qquad (47)$$

where the choice of the parameter α governs the way resources are shared among users, and its role shall be discussed later at the end of Section VI.

B. Network Utility Maximization

To account for services with asymmetric requirements, consider different DL and UL utility functions, denoted by $U_{\mathrm{DL},m}$ and $U_{\mathrm{UL},m}$ respectively for the mth user. Let $\mathbf{R}_{\mathrm{DL}}(t), \mathbf{R}_{\mathrm{UL}}(t) \in \mathbb{R}_+^{\mathrm{M}\times 1}$ denote the vectors corresponding to the vertical stacking of DL and UL per-user throughputs at the beginning of the tth frame, respectively. As user throughput varies with time, so does user utility. A global snapshot is given by the vectors $\mathbf{U}_{\mathrm{DL}}(t), \mathbf{U}_{\mathrm{UL}}(t) \in \mathbb{R}_+^{\mathrm{M}\times 1}$, where

$$[\mathbf{U}_{\mathrm{DL}}(t)]_m = U_{\mathrm{DL},m}\left(R_{\mathrm{DL},m}(t)\right) \qquad (48)$$

and a similar expression holds for UL utilities. Using (48), we define network utility as any concave non-decreasing function of the user utilities $\mathrm{NU}(t) \equiv \mathrm{NU}(\mathbf{U}_{\mathrm{UL}}(t), \mathbf{U}_{\mathrm{DL}}(t))$. It provides a cellwide aggregate indicator rating the goodness of the scheduling and resource allocation strategy carried out at the BS as far as satisfaction of all the users of the cell is concerned. For instance, we could take a maxmin approach and set network utility as the minimum among all the users' satisfaction in either UL or DL directions, i.e.,

$$\mathrm{NU}(t) = \min_{1 \leq m \leq \mathrm{M}} \left\{ \min \left\{ [\mathbf{U}_{\mathrm{UL}}(t)]_m, [\mathbf{U}_{\mathrm{DL}}(t)]_m \right\} \right\}. \qquad (49)$$

Thanks to the concavity of each user utility on the throughput and the fact that NU is a concave non-decreasing function of the utilities, it follows from the convexity properties of composite functions [37] that (49) is a concave function of $(\mathbf{R}_{\mathrm{UL}}(t), \mathbf{R}_{\mathrm{DL}}(t))$. This is an important property since concavity of network utility is necessary for obtaining global optimal allocation strategies in polynomial time. Alternatively, if there is no pressure to focus on the utility achieved by the worst user, we can take a simpler choice and set network utility as the sum of all the user's utilities

$$\mathrm{NU}(t) = \mathbf{1}_{\mathrm{M}}^T \left(\mathbf{U}_{\mathrm{UL}}(t) + \mathbf{U}_{\mathrm{DL}}(t)\right). \qquad (50)$$

When (50) is used in conjunction with (47), the parameter α is said to enable α-fairness [38]. Fairness is a wide concept which refers to the fact of not penalizing some users arbitrarily, and by tuning α from 0 to $+\infty$ the network planner is given a tool to easily switch between popular fair schemes. While $\alpha = 0$ yields utilities equal to throughputs and therefore the objective becomes maximizing the cell throughput, $\alpha = 1$ yields proportional fairness [24], and as $\alpha \to +\infty$ the network operation tends to apply the maxmin criterion to the user throughputs.

1) Optimal Strategy: For a given CSI, the task of the BS is to maximize network utility until the CSI becomes outdated and a new one is received (this period spans D frames). Afterwards, the following CSI update triggers another network utility maximization procedure for the subsequent D frames, and so on. Expressed succinctly, the optimal strategy for a given CSI is the solution to the following optimization problem[9]

$$\underset{\substack{\{\boldsymbol{\tau}_i, \mathbf{r}_{\mathrm{UL}}(t+i), \mathbf{r}_{\mathrm{DL}}(t+i)\}_{i=0}^{D-1} \\ \{\mathbf{U}_{\mathrm{UL}}(t+i), \mathbf{U}_{\mathrm{DL}}(t+i)\}_{i=1}^{D}}}{\text{maximize}} \quad \{\mathrm{NU}(t+1), \mathrm{NU}(t+2), \ldots,$$

$$\mathrm{NU}(t+D)\} \qquad (51)$$

subject to

$$[\mathbf{U}_{\mathrm{DL}}(t+i)]_m \leq U_{\mathrm{DL},m}\left(\lambda^i R_{\mathrm{DL},m}(t) + (1-\lambda) \right.$$
$$\left. \times \sum_{j=0}^{i-1} \lambda^{i-1-j} r_{\mathrm{DL},m}(t+j) \right) \qquad (52)$$

$$[\mathbf{U}_{\mathrm{UL}}(t+i)]_m \leq U_{\mathrm{UL},m}\left(\lambda^i R_{\mathrm{UL},m}(t) + (1-\lambda) \right.$$
$$\left. \times \sum_{j=0}^{i-1} \lambda^{i-1-j} r_{\mathrm{UL},m}(t+j) \right) \qquad (53)$$

$$\mathbf{r}_{\mathrm{UL}}(t+i) \in \mathcal{R}_{\mathrm{UL}}(t; \boldsymbol{\tau}_i) \qquad (54)$$
$$\mathbf{r}_{\mathrm{DL}}(t+i) \in \mathcal{R}_{\mathrm{DL}}(t; \boldsymbol{\tau}_i) \qquad (55)$$
$$\mathbf{1}_4^T \boldsymbol{\tau}_i = 1, \boldsymbol{\tau}_i \geq \mathbf{0}_4 \qquad (56)$$

where (52) and (53) apply for $1 \leq m \leq \mathrm{M}$ and $1 \leq i \leq D$, and (54)–(56) for $0 \leq i \leq D - 1$. Note that in (51)–(56), we have made implicit the resource allocation optimization with the use of the instantaneous achievable rate regions $\mathcal{R}_{\mathrm{UL}}(t; \boldsymbol{\tau}_i)$ and $\mathcal{R}_{\mathrm{DL}}(t; \boldsymbol{\tau}_i)$.

Determining the best rate allocation for the D frames under consideration amounts to solving the multiobjective optimization problem (51)–(56). Multiobjective problems do not usually have unique optimal solutions, and one usually selects one solution from the set of Pareto optimal solutions[10] according to some priorization of the objectives in conflict in the problem. Since network utility represents *cellwide* quality of service, our approach will be to provide the largest possible network utility in each of the frames under optimization indistinctly. Hence, we will first aim at maximizing the minimum network utility during

---

[9]Note that we have omitted the dependence of network utility on each of the user utilities in (51) for the sake of simplicity.

[10]Some resource allocation achieving $\{\mathrm{NU}(t+1), \mathrm{NU}(t+2), \ldots, \mathrm{NU}(t+D)\}$ belongs to the Pareto optimal set if for any other allocation achieving $\{\mathrm{NU}'(t+1), \mathrm{NU}'(t+2), \ldots, \mathrm{NU}'(t+D)\}$ it will never happen that $\mathrm{NU}'(t+i) \geq \mathrm{NU}(t+i)$ for *all* $1 \leq i \leq D$ and $\mathrm{NU}'(t+i) > \mathrm{NU}(t+i)$ for *some* $1 \leq i \leq D$.

D frames, then maximize the second smallest network utility with no penalty to the previous one, and so on. Under this criterion, one *global optimal* solution[11] can be iteratively computed using Algorithm 1.

Algorithm 1: Global maximization of network utility

1: Initializations: $\mathcal{S} = \emptyset$, $\mathrm{NU}_{\min}(t+i) = -\infty$ for $1 \leq i \leq D$.

2: **while** $|\mathcal{S}| < D$ **do**

3:   Solve

$$\underset{\substack{\{\boldsymbol{\tau}_i, \mathbf{r}_{\mathrm{UL}}(t+i), \mathbf{r}_{\mathrm{DL}}(t+i)\}_{i=0}^{D-1} \\ \{\mathbf{U}_{\mathrm{UL}}(t+i), \mathbf{U}_{\mathrm{DL}}(t+i)\}_{i=1}^{D}}}{\text{maximize}} \quad \underset{i \in \{1,2,\ldots,D\} \setminus \mathcal{S}}{\min} \{\mathrm{NU}(t+i)\} \quad (57)$$

subject to constraints (52)–(56)   (58)

$$\mathrm{NU}(t+i) \geq \mathrm{NU}_{\min}(t+i) \quad \forall i \in \mathcal{S} \quad (59)$$

4:   Compute $i_{\min} = \arg \underset{i \in \{1,2,\ldots,D\} \setminus \mathcal{S}}{\min} \{\mathrm{NU}^\star(t+i)\}$ and update $\mathcal{S} = \mathcal{S} \cup i_{\min}$, $\mathrm{NU}_{\min}(t+i_{\min}) = \mathrm{NU}^\star(t+i_{\min})$.

5: **end while**

6; Use the optimal resource allocation to compute the UL and DL exact achievable rates (2): $\{\mathbf{r}_{\mathrm{UL}}^\star(t+i), \mathbf{r}_{\mathrm{DL}}^\star(t+i)\}_{i=0}^{D-1}$.

7: Update throughputs for $1 \leq m \leq \mathsf{M}$, $1 \leq i \leq D$:

$$R_{\mathrm{UL},m}(t+i) = \lambda^D R_{\mathrm{UL},m}(t)$$
$$+ (1-\lambda) \sum_{j=0}^{i-1} \lambda^{i-1-j} r_{\mathrm{UL},m}^\star(t+j) \quad (60)$$
$$R_{\mathrm{DL},m}(t+i) = \lambda^D R_{\mathrm{DL},m}(t)$$
$$+ (1-\lambda) \sum_{j=0}^{i-1} \lambda^{i-1-j} r_{\mathrm{DL},m}^\star(t+j). \quad (61)$$

Proposition 1: The solution computed by Algorithm 1 is Pareto optimal.

Proof: See the Appendix. ∎

Remark 1: Algorithm 1 is able to compute one global optimal solution in *polynomial* time. To see this, it is sufficient to show that each of the subproblems (57)–(59) are convex. We first require that the objective (57) is concave, which follows from the concavity of network utility and the fact that the minimum of concave functions is concave. Then, the left-hand side of each of the inequality constraints in (58) and (59), when rephrased as a function of some optimization variables less than or equal to zero, should be convex. This follows from the concavity of the user utility functions with respect to throughput, the fact that the throughput relates linearly to the instantaneous rates, and the convexity of the UL and DL achievable rate regions (see Section IV).

---

[11]In case more than one global optimal resource allocation solution exists, their achieved network utility values are permuted versions of some reference $\{\mathrm{NU}^\star(t+1), \mathrm{NU}^\star(t+2), \ldots, \mathrm{NU}^\star(t+D)\}$ (see the proof of Proposition 1 in the Appendix).

Remark 2: In each of the problems (57)–(59), a threefold optimization in each of the frames under consideration is performed: first, the frame formats (relative durations of each relay-assisted transmission subphase for UL and DL) as described by the corresponding four-dimensional vectors $\{\boldsymbol{\tau}_i\}_{i=0}^{D-1}$; second, the allocated instantaneous rates, which implicitly account for user scheduling (note that $r_{\mathrm{DL},m}(t+i) = 0$ implies that the mth user shall not receive any DL data in the $(t+i)$th frame); and third, the allocated resources (bandwidth and transmit power), which are implicit in the definition of the DL and UL achievable rate regions (54), (55) as defined in (24) and (38).

Remark 3: In order to pose the maximization of network utility as a series of *convex* optimization problems, we have resorted to the concave lower bounds on the achievable rates derived in Section III-B. However, the throughput updates are performed evaluating the *exact* ergodic rates (2) achieved by the optimal resource allocation, and not their lower bounds (13). As for Rayleigh fading, we resort to [32] for their computation.

2) Reduced-Complexity Suboptimal Strategies: Although Algorithm 1 provides the best network strategy from a network utility point of view, its computational load may become prohibitive in systems either serving a large number of users per cell (large M) or refreshing the CSI slowly with respect to the frame duration (large D). To see this, consider the fact that D convex problems of $(14\mathsf{M}+4)D$ variables each (where we have considered utilities, rates, frame formats, power allocations, and bandwidth allocations) are involved in each optimization instance. Hence, two directions may be taken to cut down complexity: either reduce the optimization window D or the number of users M to be scheduled.

How to deal with the first one is immediate: simply replace D by D' in Algorithm 1 such that D' divides D (this is required to avoid optimization windows requiring some CSI not received yet). The second direction requires the implementation of a time-domain scheduler on top of Algorithm 1 such that only a subset of the M MSs are selected for network utility maximization. We choose to select the $\mathsf{M}' < \mathsf{M}$ users that would have the smallest utilities at the end of the optimization window if not scheduled. Despite this may have an impact on final performance since scheduling decisions are no longer optimal, time-domain prescheduling renders crucial in scenarios with a large number of users. On top of the aforementioned complexity issues, in practice, the frame structure needs to be signalled to the MSs, and this represents and additional overhead. If all users are allowed to transmit and/or receive in the same frame, the average quantity of the allocated resource per user and frame may go down below practical operational values while this signalling overhead may increase significantly and thus hamper network utility.

We benchmark the global optimal solution of Algorithm 1 against the suboptimal strategy which takes $D' = 1$ and attempts to maximize network utility *sequentially* in a frame-by-frame basis. Thus, Algorithm 2 is run at the beginning of each frame, which allocates resources among the subset of M' users selected by a time-domain prescheduler. This way, we are able to quantify the performance loss of sequential optimization versus global optimization.

TABLE I
PHYSICAL LAYER SETUP OF THE SIMULATED SCENARIO

B	10 MHz
T	25 ms
$p_{\text{BS}}^{\max}$, $p_{\text{RS}}^{\max}$, and $p_{\text{BS}}^{\max}$	33, 30, and 24 dBm
G_{BS}, G_{RS}, and G_{BS}	10.6, 5, and -1 dB
F_{BS}, F_{RS}, and F_{BS}	4, 4, and 7 dB
n_{BS}, n_{RS}, and n_{MS}	2, 2, and 1
BS, RS, and MS heights	15, 5, and 0 m
N_0	-114 dBm/MHz
Γ	4 dB
λ	0.95

Algorithm 2: Sequential maximization of network utility

1: Solve

$$\underset{\substack{\boldsymbol{\tau}, \mathbf{r}_{\text{UL}}(t), \mathbf{r}_{\text{DL}}(t), \\ \mathbf{U}_{\text{UL}}(t+1), \mathbf{U}_{\text{DL}}(t+1)}}{\text{maximize}} \quad \text{NU}(t+1) \tag{62}$$

subject to

$$[\mathbf{U}_{\text{DL}}(t+1)]_m$$
$$\leq U_{\text{DL},m}\left(\lambda R_{\text{DL},m}(t) + (1-\lambda)r_{\text{DL},m}(t)\right) \tag{63}$$

$$[\mathbf{U}_{\text{UL}}(t+1)]_m$$
$$\leq U_{\text{UL},m}\left(\lambda R_{\text{UL},m}(t) + (1-\lambda)r_{\text{UL},m}(t)\right) \tag{64}$$

$$\mathbf{r}_{\text{UL}}(t) \in \mathcal{R}_{\text{UL}}(t; \boldsymbol{\tau}) \tag{65}$$

$$\mathbf{r}_{\text{DL}}(t) \in \mathcal{R}_{\text{DL}}(t; \boldsymbol{\tau}) \tag{66}$$

$$\mathbf{1}_4^T \boldsymbol{\tau} = 1, \boldsymbol{\tau} \geq \mathbf{0}_4. \tag{67}$$

2: Use the optimal resource allocation to compute the UL and DL exact achievable rates (2): $\mathbf{r}_{\text{UL}}^\star(t), \mathbf{r}_{\text{DL}}^\star(t)$.

3: Update throughputs for $1 \leq m \leq \mathsf{M}$:

$$R_{\text{UL},m}(t+1) = \lambda R_{\text{UL},m}(t) + (1-\lambda)r_{\text{UL},m}^\star(t) \tag{68}$$
$$R_{\text{DL},m}(t+1) = \lambda R_{\text{DL},m}(t) + (1-\lambda)r_{\text{DL},m}^\star(t). \tag{69}$$

VI. SIMULATION RESULTS

We focus on two different scenarios sharing the same target cell spectral efficiency but having different user population sizes. In either case, we simulate a circular cell of 500 m radius, with $\mathsf{R} = 5$ relays uniformly spaced along a circle at 375 m from the BS, which is located in the cell center. We assume that the MS–RS links are in line-of-sight and have path loss exponent 2.6, while the rest of links (BS–MS and MS–RS) are in non line-of-sight with path loss exponent 4.05. All links are hampered by Rayleigh fading. See Table I for a complete list of values of the rest of physical parameters involved.

All the users of the cell are mobile. If $\mathbf{x}_m(t) \in \mathbb{R}_+^2$ denotes the position of the mth MS at the beginning of the tth frame in Cartesian coordinates, then

$$\mathbf{x}_m(t+1) = \mathbf{x}_m(t) + v_m T \begin{bmatrix} \cos(\varphi(t)) \\ \sin(\varphi(t)) \end{bmatrix} \tag{70}$$

where v_m is its speed (assumed constant) and $\varphi(t)$ is an AR(1) stochastic process describing its direction

$$\varphi(t+1) = 0.9\varphi(t) + 0.02\psi_t \tag{71}$$

where $\{\psi_t\}$ are i.i.d. uniform random variables on $-[\pi, \pi)$. Whenever a MS happens to exceed the limits of the cell, it is forced to bounce on the cell edge by reversing its instantaneous direction in order to keep constant the total number of users. As a simplifying assumption we set the same speed of $v = 3\,\text{kmph}$ for all the MSs, and consider a feedback update rate of 100 ms, i.e., $D = 4$. Each time CSI is refreshed, each MS is attached to the RS towards which the path loss is the smallest. The connectivity matrix $\mathbf{L}$ is updated accordingly.

With this setup, we first simulate an scenario consisting of $\mathsf{M} = 6$ best-effort users (44) of gold, silver, and bronze QoS classes. Gold users experience 0.9 utility when they are given 30 Mb/s (DL) and 6 Mb/s (UL) throughput. Silver users have the same utility level when served 20 Mb/s (DL) and 4 Mb/s (UL) throughput. Finally, bronze users require 10 Mb/s (DL) and 2 Mb/s (UL) throughput for 0.9 utility. There are two users of each QoS class and, under a maxmin choice (49), if a *network* utility of 0.9 was realized, the cell spectral efficiency would amount to 14.4 b/s/Hz.

Fig. 4 shows the deployment layout and compares the network utility achieved by global and sequential optimization. To quantify separately the performance gains provided by the presence of relays from those achieved by the optimization approach itself, we benchmark the global and the sequential optimization algorithms against their counterparts without RSs. That is, we also simulate resource allocation where the relay-transmit phase is always forced to have zero duration. To make this comparison fair, we increase the transmit power constraint at the BS and the MSs such that, frame by frame, the total UL and DL power is equal with and without RSs in both optimization strategies. Since the number of users is relatively small, there is no need for a time-domain prescheduler.

As a general trend, global optimization dominates sequential optimization in the long term, although at the beginning the opposite holds. This is because in the first frames, the resource allocation of sequential optimization benefits from evaluating the actual ergodic rates often (frame by frame) as compared to global optimization, where this evaluation is carried out every group of D frames. This has a positive impact on user throughput, although this advantage vanishes quickly as confirmed also by Fig. 5, where the per-user throughput achieved by global optimization dominates that of sequential optimization in less than 2 s of network operation (80 frames). Interestingly, sequential and global optimization perform almost undistinguishably without relays: as the ergodic capacity lower bounds are very tight in this setup (see Fig. 2), the previous effect does not apply. In any case, the target spectral efficiency of 14.4 b/s/Hz is achieved only with relaying infrastructure, as the steady-state performance without RSs falls roughly 25% below QoS targets. Indeed, as showed in Table II the price to pay is an increase in computation complexity.

Next, we focus on a practical scenario with the same target spectral efficiency at 0.9 network utility as before, adopting the

TABLE II
RELATIVE EXECUTION TIMES PER TRANSMISSION FRAME OF THE MATLAB IMPLEMENTATIONS OF ALL THE RESOURCE ALLOCATION STRATEGIES

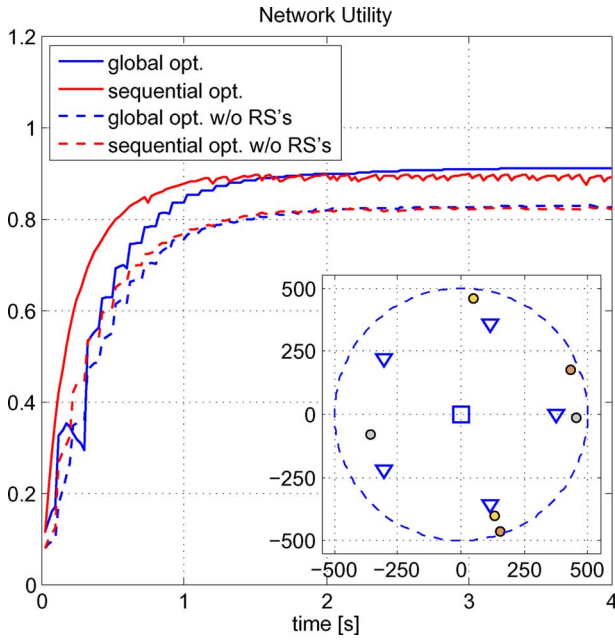| Seq. Opt. w/o RS's	Global Opt. w/o RS's	Seq. Opt. w/ RS's	Global Opt. w/ RS's
1	1	2.6	33

Fig. 4. Network utility achieved by global (blue) and sequential (red) optimization, with (solid) and without (dashed) relaying infrastructure, and deployment layout.
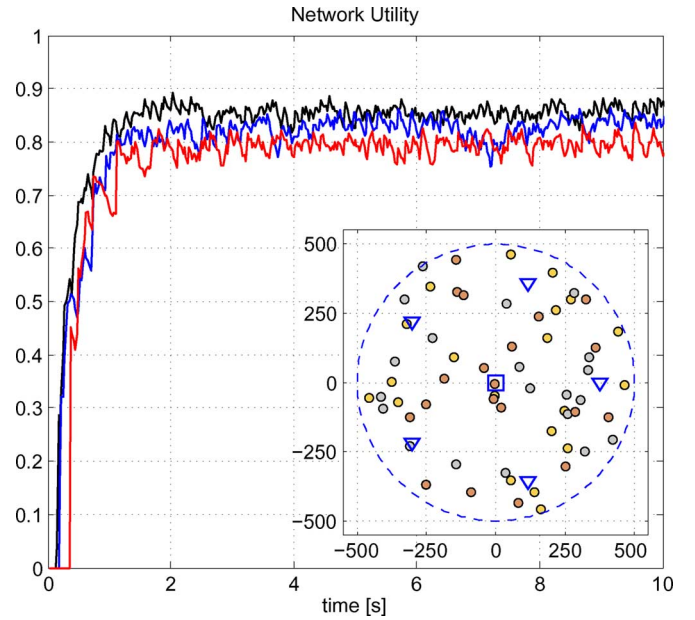

Fig. 6. Network utility achieved by sequential optimization with time-domain prescheduling allowing a maximum of 12 (black), 8 (blue), and 4 (red) users per frame, and deployment layout.
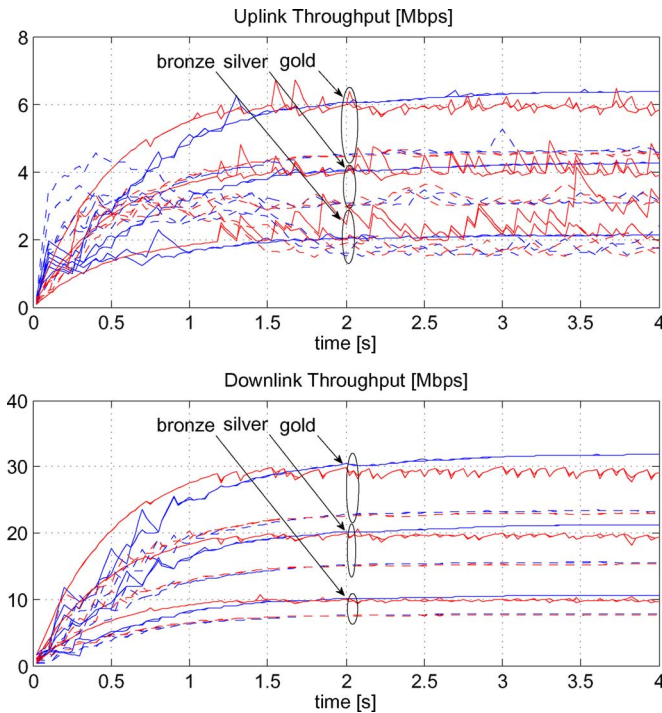

Fig. 5. Per-user served throughput of global (blue) and sequential (red) optimization, with (solid) and without (dashed) relaying infrastructure.

which split into 20 users per QoS class. Now, at 0.9 user utility gold users require 3/0.6 Mb/s (DL/UL) throughput, silver users 2/0.4 Mb/s (DL/UL) throughput, and bronze users 1/0.2 Mb/s (DL/UL) throughput. With this system size, global optimization renders unfeasible and we restrict our attention to sequential optimization with time-domain prescheduling. In particular, we study the performance degradation as a function of M', the maximum number of users per frame. In this respect, Fig. 6 shows the deployment layout and the network utility achieved with $M' = 12$, 8, and 4. Clearly, the larger M', the larger network utility but also the algorithm complexity and signalling overhead. Assuming negligible this last effect for the range of values of M' studied,[12] the steady-state network utility loss between $M' = 12$ and $M' = 4$ is on the order of 0.1.

In Fig. 7 we show the per-user throughput averaged per QoS class, to show that, although the general trend is similar for each value of M', the performance degradation when $M' = 4$ is due to the fact that gold users are served far below their 0.9 target while bronze users are satiated more than necessary. As M' decreases, the average number of idle frames between transmissions for a given user increases, but the instantaneous rate per user in an active frame increases. The conclusion from Fig. 7 is that if M'

maxmin utility criterion (49) again, but now serving 10 times more users with rate requirements reduced to 10% of the previous. Thus, $M = 60$ best-effort MSs are present in the cell

---

[12]For each direction (UL and DL), each user should be signalled about the transmission rate, the fractional bandwidth allocation, the fractional power allocation, and the duration of the protocol subphases. Assuming that each of these parameters is quantized to n_b bits, the throughput penalty per user and direction is $n_b(6 + 2/M')$ which decreases with M' since the subphase durations are common. Hence, although the global signalling overhead increases in M', the per-user penalty decreases. In particular for a practical value of $n_b = 5$ bits, this penalty is on the order of 1 kb/s and, hence, negligible.
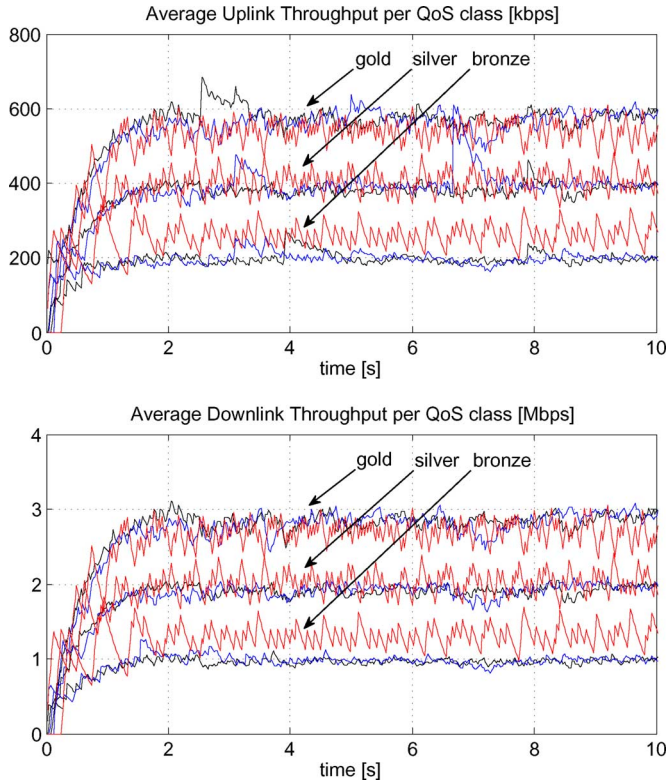
Fig. 7. Average per-user served throughput per QoS class achieved by sequential optimization with time-domain prescheduling allowing a maximum of 12 (black), 8 (blue), and 4 (red) users per frame.
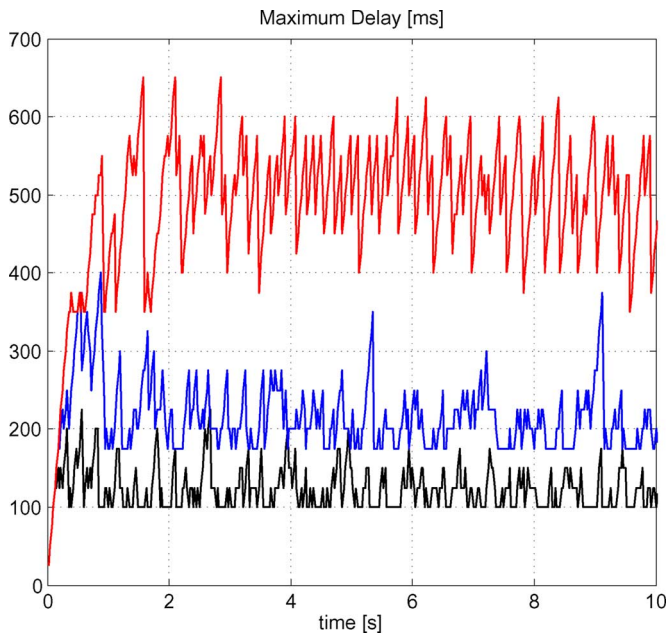

Fig. 8. Maximum user delay (number of frames idle) achieved by sequential optimization with time-domain prescheduling allowing a maximum of 12 (black), 8 (blue), and 4 (red) users per frame.

is too low the global effect is negative. Finally, Fig. 8 addresses the maximum per-user delay in each setup, which is shown to be roughly proportional to $1/\mathsf{M}'$ in average.
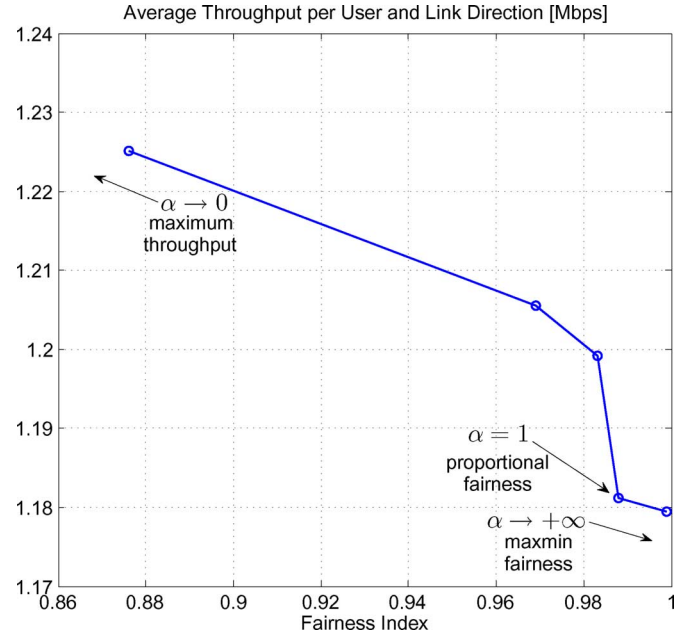

Fig. 9. Average steady-state per-user and link-direction throughput versus fairness index. The corresponding values of α are 0.25, 0.50, 0.75, 1, and 5.

So far, we have only explored maxmin network utility and QoS-oriented utility functions. To conclude this section, however, we shall modify the previous system with a total of $\mathsf{M} = 60$ users and $\mathsf{M}' = 4$ users at most per frame and explore the inherent tradeoff between fairness and throughput when other network utility functions are used. In particular, we now consider the situation where the UL and DL utility functions of every user are given by (47) and network utility is sum utility (50). Thus, by changing the parameter α we have a way of trading fairness and throughput. We capture this relation in Fig. 9, where we focus on the average steady-state per-user and link direction throughput (UL and DL directions are averaged together as their utilities are the same). It is plotted against Jain's fairness index [39] in each direction for a given cell deployment. This index rates the degree of fairness incurred in serving n competing flows with a real number between $1/n$ (worst case: one user gets it all) and 1 (best case: resources are equally shared). Fig. 9 confirms that a simultaneous increase in both fairness and throughput cannot be achieved and quantifies the explicit tradeoff. Notice that the time-domain prescheduler prevents the fairness index to fall below acceptable levels.

VII. CONCLUSION

This paper concentrates on the performance characterization of a relay-assisted network deployment under practical constraints. In particular, terminals are half-duplex MIMO and path loss is the only CSI assumed to be known at the transmitters. Network performance is evaluated in terms of ergodic achievable rates by the development of novel lower bounds. These bounds are employed to derive two efficient algorithms for resource allocation optimization under heterogeneous QoS requirements. The first one provides one Pareto optimal solution whereas the second one performs a simpler frame by frame optimization by means of a sequential algorithm. The

performance of both schemes has been evaluated showing that, whenever global optimization can be afforded, significant performance gains can be achieved with respect to sequential optimization. When systems dimensions are large, however, the complexity of sequential network optimization can be tuned by using time-domain prescheduling.

The proposed network resource optimization schemes could be generalized in at least two ways. First, by incorporating outage events into system design in those scenarios where the number of tones is limited and moderate. Second, by considering a multi-cell configuration allowing the incorporation of a prescribed maximum inter-cell interference as an additional constraint.

APPENDIX
PROOF OF PROPOSITION 1

Let $\{NU^\star(t+1), NU^\star(t+2), \ldots, NU^\star(t+D)\}$ denote the network utility achieved by the solution of Algorithm 1, and define $i_1, i_2, \ldots, i_D$ such that $NU^\star(t+i_1) \leq NU^\star(t+i_2) \leq \ldots \leq NU^\star(t+i_D)$. Assume it is *not* Pareto optimal. Hence, there exists at least another allocation achieving $\{NU(t+1), NU(t+2), \ldots, NU(t+D)\}$ and integers $j_1, j_2, \ldots, j_D$ such that $NU(t+j_1) \leq NU(t+j_2) \leq \ldots \leq NU(t+j_D)$,

$$NU(t+i) \geq NU^\star(t+i), \quad 1 \leq i \leq D, \tag{72}$$

$$NU(t+i) > NU^\star(t+i), \quad \text{for some } i. \tag{73}$$

Then,

$$NU^\star(t+i_1) \overset{(a)}{\geq} NU(t+j_1) \overset{(b)}{\geq} NU^\star(t+j_1) \overset{(c)}{\geq} NU^\star(t+i_1) \tag{74}$$

where (a) follows by construction of the solution of Algorithm 1, (b) from (72), and (c) from the definition of $\{i_n\}$. Equation (74) implies that $NU^\star(t+i_1) = NU(t+j_1)$. Two cases arise now:

- Case $i_1 \neq j_1$:

$$NU^\star(t+j_1) \overset{(a)}{\geq} NU^\star(t+i_1) \overset{(b)}{=} NU(t+j_1) \overset{(c)}{\geq} NU^\star(t+j_1) \tag{75}$$

where (a) follows from the definition of $\{i_n\}$, (b) is a consequence of (74), and (c) follows from (72). Then, $NU^\star(t+j_1) = NU(t+j_1)$. Without loss of generality we can set $i_2 = j_1, j_2 = i_1$ and obtain $NU^\star(t+i_2) = NU(t+j_2)$.

- Case $i_1 = j_1$:

$$NU^\star(t+i_2) \overset{(a)}{\geq} NU(t+j_2) \overset{(b)}{\geq} NU^\star(t+j_2) \overset{(c)}{\geq} NU^\star(t+i_2) \tag{76}$$

where (a) follows by construction of the solution of Algorithm 1, (b) is derived from (72), and (c) follows from the fact that $j_2 \neq j_1 = i_1$ and the definition of $\{i_n\}$. Expression (76) also implies $NU^\star(t+i_2) = NU(t+j_2)$.

Proceeding similarly, we can iteratively show that

$$NU^\star(t+i_n) = NU(t+j_n), \quad 1 \leq n \leq D \tag{77}$$

which implies that the network utility values of both strategies are either equal or related by an arbitrary permutation. In either

case, this contradicts (72), (73), hence implying that the solution to Algorithm 1 is Pareto optimal. $\square$

REFERENCES

[1] T. M. Cover and A. A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inf. Theory*, vol. 25, pp. 572–584, Sep. 1979.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, Jul. 2006.

[3] R. Nabar, H. Bölcskei, and F. Kneubühler, "Fading relay channels: Performance limits and space-time signal design," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 1099–1109, Aug. 2004.

[4] B. Wang, J. Zhang, and A. Host-Madsen, "On the capacity of MIMO relay channels," *IEEE Trans. Inf. Theory*, vol. 51, pp. 29–43, Jan. 2005.

[5] H. Ochiani, P. Mitran, and V. Tarokh, "Variable rate two phase collaborative communication protocols for wireless networks," *IEEE Trans. Inf. Theory*, vol. 52, pp. 4299–4313, Sep. 2006.

[6] A. Høst-Madsen and J. Zhang, "Capacity bounds and power allocation for wireless relay channels," *IEEE Trans. Inf. Theory*, vol. 51, pp. 2020–2040, Jun. 2005.

[7] M. Yu and J. Li, "Is amplify-and-forward practically better than decode-and-forward or vice versa?," presented at the IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Philadelphia, PA, Mar. 2005.

[8] M. C. Valenti and B. Zhao, "Distributed turbo codes: Towards the capacity of the relay channel," presented at the IEEE Vehicular Technology Conf.—Fall, Orlando, FL, Oct. 2003.

[9] M. Dohler, A. Gkelias, and H. Aghvami, "A resource allocation strategy for distributed MIMO multi-hop communication systems," *IEEE Commun. Lett.*, pp. 99–101, Feb. 2004.

[10] E. Biglieri and G. Taricco, *Transmission and Reception With Multiple Antennas: Theoretical Foundations*. Delft, The Netherlands: Now Publishers, 2004.

[11] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[12] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor, *MIMO Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[13] Y.-J. Chang, F.-T. Chien, and C.-C. J. Kuo, "Cross-layer QoS analysis of opportunistic OFDM-TDMA and OFDMA networks," *IEEE J. Sel. Areas Commun.*, vol. 25, pp. 657–666, May 2007.

[14] *IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*, IEEE Std. 802.16e-2005, Tech. Rep. IEEE Standards Dept., Dec. 2005, New York.

[15] IEEE 802.16 Relay Task Group, "Multi-hop relay system evaluation methodology (channel model and performance metric)," IEEE 802.16j-06/013r2, Nov. 2006.

[16] "IEEE 802.16 Broadband Wireless Access Group, IEEE 802.16 m system requirements," IEEE 802.16m-07/002r4, Oct. 2007.

[17] W. Yu and R. Lui, "Dual methods for non-convex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, pp. 1310–1322, Jul. 2006.

[18] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, pp. 1747–1758, Oct. 1999.

[19] C. Bae and D.-H. Cho, "Adaptive resource allocation based on channel information in multihop OFDM systems," presented at the IEEE Vehicular Technology Conf. (VTC)—Fall, Montreal, QC, Canada, Sep. 2006.

[20] K.-D. Lee and V. C. M. Leung, "Fair allocation of subcarrier and power in an OFDMA wireless mesh network," *IEEE J. Sel. Areas Commun.*, vol. 24, pp. 2051–2060, Nov. 2006.

[21] T. C.-Y. Ng and W. Yu, "Joint optimization of relay strategies and resource allocations in cooperative cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 25, pp. 328–339, Feb. 2007.

[22] *Evolved Universal Terrestrial Radio Access (E-UTRA), Long Term Evolution (LTE) Physical Layer, General Description*, 3GPP TS 36.201, 3GPP TS 36.201.

[23] K. H. Teo, Z. Tao, and J. Zhang, "The mobile broadband WiMAX standard," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 144–148, Sep. 2007.

[24] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: Shadow prices, proportional fairness, and stability," *J. Oper. Res. Soc.*, vol. 49, pp. 237–252, Mar. 1998.

[25] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, pp. 1452–1463, Aug. 2006.

[26] D. P. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: Framework and applications," *IEEE Trans. Autom. Control*, vol. 52, pp. 2254–2269, Dec. 2007.

[27] Y. Xue, B. Li, and K. Nahrstedt, "Optimal resource allocation in wireless ad hoc networks: A price-based approach," *IEEE Trans. Mobile Comput.*, vol. 5, pp. 347–364, Apr. 2006.

[28] Y. Liang, V. V. Veeravalli, and H. V. Poor, "Resource allocation for wireless fading relay channels: max-min solution," *IEEE Trans. Inf. Theory*, vol. 53, pp. 3432–3453, Oct. 2007.

[29] M. Dohler, A. Gkelias, and H. Aghvami, "Resource allocation for FDMA-based regenerative multihop links," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1989–1993, Nov. 2004.

[30] E. Calvo, J. Vidal, and J. R. Fonollosa, "Resource allocation in multihop OFDMA broadcast networks," presented at the IEEE SPAWC, Helsinki, Finland, Jun. 2007.

[31] H. Shin and J. H. Lee, "Capacity of multiple-antenna fading channels: Spatial fading correlation, double scattering, and keyhole," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2636–2647, Oct. 2003.

[32] M. Kießling, "Unifying analysis of ergodic MIMO capacity in correlated Rayleigh fading environments," *Eur. Trans. Telecommun.*, vol. 16, pp. 17–35, Jan. 2005.

[33] M. Dohler and H. Aghvami, "On the approximation of MIMO capacity," *IEEE. Trans. Wireless Commun.*, vol. 4, pp. 30–34, Jan. 2005.

[34] M. Fiedler, "Bounds for the determinant of the sum of Hermitian matrices," *Proc. Amer. Math. Soc.*, vol. 30, pp. 27–31, Sep. 1971.

[35] A. Lapidoth and S. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat fading channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2426–2467, Oct. 2003.

[36] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed.   New York: Academic, 2000.

[37] S. P. Boyd and L. Vandenberghe, *Convex Optimization*.   Cambridge, U.K.: Cambridge Univ. Press, 2004.

[38] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, pp. 556–567, Oct. 2000.

[39] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared systems," DEC Res. Rep.TR-301, 1984.

Josep Vidal (M'91) received the Telecommunication Engineering and the Ph.D. degrees from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain.

From 1989 to 1990, he was an Associate Researcher with the LTS at the Ecole Polytechnique de Lausanne. He was awarded Premio Extraordinario de Doctorado in 1996. Since 1993 he has been teaching different topics in telecommunication engineering at UPC. His current research interests are in statistical signal processing, information, and communication theory, where he has authored more than 100 journal and conference papers. Since 2000 has led UPC participation in the EC-funded projects, SATURN, ROMANTIK, FIREWORKS and ROCKET, and served as project coordinator in some of them. He has held research appointments with INP Toulouse and the University of Hawaii.

Javier Rodríguez Fonollosa (S'90–M'92–SM'98) received the Ph.D. degree in electrical and computer engineering at Northeastern University, Boston, MA, in 1992.

In 1993, he joined the Department of Signal Theory and Communications of the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, where he became Associate Professor in 1996, Professor in 2003, and Department Head in 2006. He is the author of more than 100 papers in the area of signal processing and communications. In 1995, he led UPC's participation in the European Commission funded ACTS Mobile projects TSUNAMI(II) and SUNBEAM that included the analysis of adaptive antennas in second- and third-generation cellular mobile communication systems. In January 2000, he was appointed technical and project coordinator of the IST project METRA dedicated to the introduction of multiantenna terminals in UMTS. This project continued until 2003 under the name of I-METRA, looking into more advanced systems and systems beyond 3G. From January 2006 to December 2008, he coordinated the Sixth Framework Programme IST project SURFACE, which evaluated the performance of a generalized air interface with self-configuration capabilities. Since October 2006 he has been project coordinator of the five-year Type C-Consolider project Fundamental bounds in Network Information Theory of the National Research Plan of Spain. Also since October 2008, he has been project coordinator of the five-year Consolider-Ingenio project COMONSENS on Foundations and Methodologies for Future Communication and Sensor Networks. His research interests include many different aspects of statistical signal processing for communications and information theory.

Dr. Rodríguez Fonollosa was Co-Chairman and organizer of the IEEE Signal Processing/ATHOS Workshop on Higher-Order Statistics held in Begur, Girona, Spain, in June 1995 and September 2001 and of the IST Mobile Communications Summit held in Sitges, Barcelona, Spain, in 2001. He was elected member of the Signal Processing for Communications (SPCOM) Technical Committee of the IEEE Signal Processing Society in January 1999. Since May 2005, he has been a member of the Editorial Board of the EURASIP *Signal Processing Journal*.

Eduard Calvo (S'02) received the M.S. (with highest hons.) and Ph.D. degrees in electrical engineering from the Technical University of Catalonia (UPC), Spain, in 2004 and 2009, respectively.

In 2009, he joined IESE Business School, Universidad de Navarra, as a Postdoctoral Fellow at the Department of Production, Technology, and Operations Management. He has participated in several European projects from the 6th and 7th Framework program. During 2004 and 2008, he held research appointments at the Massachusetts Institute of Technology (MIT) His research interests lie in the intersection of information theory, optimization theory, statistical signal processing, and their applications to wireless and underwater communication networks, and operations research.

Dr. Calvo was given the Salvà i Campillo award to the Best Young Engineer by the College of Electrical Engineers of Catalonia in 2005.