

# Robust Scheduling and Resource Allocation in the Downlink of Spatially Correlated MIMO-OFDMA Wireless Systems With Imperfect CSIT



# Robust Scheduling and Resource Allocation in the Downlink of Spatially Correlated MIMO-OFDMA Wireless Systems with Imperfect CSIT

Guillem Femenias, *Senior Member, IEEE,*
Felip Riera-Palou, *Senior Member, IEEE*, and
John S. Thompson, *Senior Member, IEEE*

**Abstract**

MIMO-OFDMA has been selected as the core physical layer access scheme for the downlink of state-of-the-art and next-generation wireless communications standards. In these systems, scheduling and resource allocation algorithms, jointly assigning transmission data rates, bandwidth and power, become crucial to optimize the resource utilization while providing support to multimedia applications with heterogeneous quality of service (QoS) requirements. To this end, the transmitter is assumed to have channel state information (CSIT) that will typically be imperfect. This paper introduces a unified analytical framework for robust channel- and queue-aware QoS-guaranteed cross-layer scheduling and resource allocation algorithms for the downlink of MIMO-OFDMA networks with imperfect CSIT. The framework is based on the statistical characterization of the SNR under imperfect CSIT, and is general enough to encompass spatial correlation effects in the Tx and Rx antenna arrays, different types of traffic, uniform and continuous power allocation, discrete and continuous rate allocation, and protocols with different amounts of channel- and queue-awareness. Simulation results using parameters drawn from the 3GPP-LTE standard demonstrate the validity and merits of the proposed robust cross-layer unified approach.

G. Femenias and F. Riera-Palou are with the Mobile Communications Group, University of the Balearic Islands (UIB), Mallorca 07122, Illes Balears (Spain).

J.S. Thompson is with the Institute for Digital Communications, University of Edinburgh, Edinburgh EH9 3JL, Scotland (UK).

# I. INTRODUCTION

Orthogonal frequency division multiple access (OFDMA), combined with multiple-input multiple-output (MIMO) strategies, has been chosen as the core physical layer access scheme for state-of-the-art and next-generation wireless communications standards such as IEEE 802.16e/m-based WiMAX systems [1] and the Third Generation Partnership Project (3GPP) networks based on the Long-Term Evolution (LTE) and LTE-Advanced (LTE-A) [2]. In the downlink of a MIMO-OFDMA system, the scheduling and resource allocation (SRA) unit at the base station (BS) obtains channel state information (CSI) from the physical (PHY) layer of all the mobile stations (MSs) in the system and collects queue state information (QSI) by observing the backlogged data in the buffers of these MSs at the data link control (DLC) layer. Based on this information, the SRA unit can then make a cross-layer SRA decision allowing a good trade-off between multiuser diversity exploitation, provision of fairness and delivering of quality-of-service (QoS) to the wide range of applications supported by emerging OFDMA-based wireless networks [3].

There are quite a number of existing works reporting optimal and suboptimal SRA algorithms taking into consideration, in a cross-layer fashion, the DLC layer bursty packet arrivals and queueing behavior jointly with the PHY layer channel conditions [4]–[11]. Most of them, however, assume that the BS has access to ideal CSI whereas, although accurate CSI is essential to achieve high spectral efficiency while providing heterogeneous QoS requirements, perfect instantaneous CSI is rarely available in practice due to estimation errors and/or feedback delay. The problem of joint SRA in the OFDMA downlink under imperfect CSI has been previously studied in several papers, notably [12]–[17]. Wang and Lau in [12], focus on a cross-layer MIMO-OFDMA design in which zero-forcing beamforming is used at the BS. Suboptimal algorithms are designed for ergodic continuous rate allocation using a Gaussian convergence approximation for the conditional outage probability and a simplification of the combinatorial search in the scheduling step. The Gaussian convergence lemma, however, is only valid when the number of users allocated to a given subcarrier is sufficiently large, a condition rarely fulfilled in real systems, and the simplified combinatorial search is only valid when the total transmit power available at the BS is very large. In [13], Wong and Evans propose a dual optimization approach to derive optimal resource allocation algorithms for ergodic continuous and discrete rate maximization assuming the availability of partial CSI. In the calculation of the expected sum rate, however, they do not take into account that capacity outages are produced when, due to

imperfect CSI, the data-rate allocated by the BS to a given MS is greater than the instantaneous channel capacity, a situation quite frequent in systems without a properly designed rate backoff function [18]. A similar approach was proposed by Awad *et al.* in [15], where the calculations of the achievable rate as a function of the imperfect CSI do not consider the packet losses induced by capacity outages. In [14], the authors also use dual decomposition techniques to solve a joint SRA optimization problem in which multiple users can time-share each of the subcarriers in the system. The optimization problem, however, is formulated using upper bounds on the capacity of the system (via Jensen's inequality) which are only applicable to capacity-based continuous rate allocation systems. Aggarwal *et al.* in [16] consider maximizing the ergodic sum-utility by using a generic probability distribution to model the imperfect CSI. The proposed solution, however, is only applicable to capacity-based continuous data-rate allocation systems and, furthermore, eventual capacity outages are modeled using far from practical bit error rate (BER) approximations. Finally, Zarakovitis *et al.* in [17] present a cross-layer design for SISO-OFDMA systems with imperfect CSI and considering data outage effects. The authors of this paper claim that goodput-based optimization provides infeasible scheduling operation under channel conditions with average to high uncertainty, an issue that we address in this paper, and propose a power-bit loading algorithm aiming at power-efficiency maximization.

In this paper, cross-layer SRA algorithms addressing some of the shortcomings of the previous approaches are proposed. These algorithms are able to optimally exploit spatially correlated MIMO-OFDMA wireless systems with imperfect channel state at the transmitter (CSIT). Our main contributions in this paper can be summarized as follows:

1) The BS has to take SRA decisions based on the availability of partial CSI at the SRA unit. In order to minimize the negative effects of a bad scheduling and/or allocation decision the SRA unit should have access to a proper mathematical model characterizing the behavior of the propagation channel under imperfect CSI. In Section III a thorough statistical characterization of the received signal-to-noise ratio (SNR) conditioned on the partial CSI available at the SRA unit is provided. The mathematical model is general enough to account for the joint effects of imperfect CSIT and MIMO channel correlation.
2) In contrast to previous works (i.e., [12]–[17]), our paper proposes a unified framework that, based on the statistical characterization of the SNR, takes into account the packet outage probability due to imperfect CSI and generalizes results presented in, for instance, [6], [8], [9], [11], [19]. To this end, Section IV introduces a framework able to account for: (i) the

existence of poor scattering conditions in the surroundings of both the transmit and receive terminals, (ii) different types of traffic, (iii) different allocation strategies, (iv) protocols with different amounts of CSI- and QSI-awareness, and (v) different utility functions measuring user's satisfaction.

3) Reinforcing the practical applicability of the proposed framework, results for the discrete data-rate schemes presented in Section V are obtained using block error probability models obtained for state-of-the-art adaptive modulation and coding (AMC) schemes standardized by the 3GPP [20]–[22].

The rest of the paper is organized as follows. Section II introduces the system model and related traffic assumptions. Section III presents a statistically robust design of the SRA unit that is based on the statistical characterization of the SNR in light of imperfect CSIT and spatial correlation. Section IV formally states the objective optimization problem whose solution is tackled using convex optimization techniques. Extensive numerical results are provided in Section V. Finally, the main outcomes of this paper are recapped in Section VI

In this paper, vectors and matrices are denoted by lower- and uppercase bold letters, respectively. The K-dimensional identity matrix is represented by $\boldsymbol{I}_K$. The symbols $\mathbb{R}_+$ and $\mathbb{C}$ serve to denote the set of non-negative real numbers and the set of complex numbers, respectively. The operator $\mathrm{tr}\{\cdot\}$ denotes the trace of a matrix whereas $\otimes$ represents the Kronecker product of two matrices. Superscripts $(\cdot)^T$ and $(\cdot)^H$ are used to denote the transpose and the conjugate transpose (hermitian) of a matrix.

## II. System Model and assumptions

Let us consider the downlink of a time-slotted MIMO-OFDMA wireless packet access network as the one depicted in Fig. 1. In this setup, a BS with a total transmit power $P_T$ and equipped with $N_T$ transmit antennas provides service to $N_m$ active MSs, indexed by the set $\mathcal{N}_m = \{1, \ldots, N_m\}$, each equipped with $N_R^{(m)}$ receive antennas.

Transmission between the BS and active MSs is organized in time-frequency resource allocation units, also known as resource blocks (RB). Each RB is formed by a slot in the time axis and a subband in the frequency axis:

- In the time axis, each RB occupies a time-slot of a fixed duration $T_s$, assumed to be less than the channel coherence time. Thus, the channel fading can be considered constant over the whole slot and it only varies from slot to slot, i.e., a slot-based block fading channel is

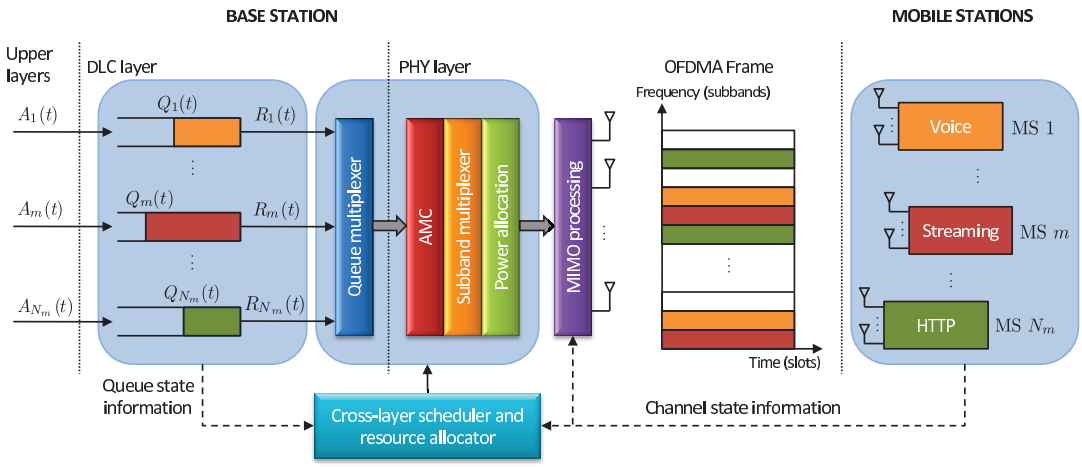


Figure 1: System model.

assumed. Each of these slots consists of a fixed number $N_o$ of orthogonal frequency-division multiplexing (OFDM) symbols of duration $T_o + T_{CP} = T_s/N_o$, where $T_{CP}$ is the cyclic prefix (CP) duration.

- Slotted transmissions take place over a bandwidth $B$, which is divided into $N_f$ orthogonal subcarriers, out of which $N_d$ are used to transmit data and $N_p$ are used to transmit pilots and to set guard frequency bands. The $N_d$ data subcarriers are divided into $N_b$ orthogonal subbands, each consisting of $N_{sc}$ adjacent subcarriers and with a bandwidth $B_b = B\,N_d/(N_f\,N_{sc})$ small enough to assume that all subcarriers in a subband experience frequency flat fading. Frequency subbands in a given slot are indexed by the set $\mathcal{N}_b = \{1, \ldots, N_b\}$.

Without loss of generality, and in order to simplify the mathematical notation of the problem, only one service data flow (also known as a connection or session) per active MS will be assumed. Depending on the traffic type, three classes of service and the associated QoS requirements and priorities must be accounted for in wireless communications[1] [24]: *Best Effort* (BE), *Non-Real-Time* (nRT), and *Real-Time* (RT). Traffic flows arriving from higher layers are buffered into the corresponding $N_m$ first-in first-out (FIFO) queues at the DLC layer. At the beginning of each scheduling time interval, based on the available joint CSI/QSI, the cross-layer SRA unit selects

[1]Using LTE terminology, real-time services would be associated to guaranteed bit rate (GBR) evolved packet system (EPS) bearers, and the non GBR bearers would be suitable for best-effort and non-real-time services [23].

some packets in the queues for transmission, which are then forwarded to the OFDM transmitter, where they are adaptively modulated and channel encoded and, based on the MIMO processing, are allocated power and subbands.

### A. Channel model

Relying on the widely used Kronecker channel model [25], the spatially-correlated MIMO channel matrix between the BS and MS $m$ on subband $b$ can be expressed as

$$\boldsymbol{H}_{b,m}(t) = \boldsymbol{R}_{Rb,m}^{1/2} \boldsymbol{\mathcal{H}}_{b,m}(t) \boldsymbol{R}_{Tb,m}^{1/2}, \tag{1}$$

where

$$\boldsymbol{\mathcal{H}}_{b,m}(t) = \begin{bmatrix} \mathcal{H}_{b,m}^{1,1}(t) & \dots & \mathcal{H}_{b,m}^{1,N_T}(t) \\ \vdots & & \vdots \\ \mathcal{H}_{b,m}^{N_R^{(m)},1}(t) & \dots & \mathcal{H}_{b,m}^{N_R^{(m)},N_T}(t) \end{bmatrix}. \tag{2}$$

is a spatially white complex valued $N_R^{(m)} \times N_T$ MIMO matrix and the matrices $\boldsymbol{R}_{Tb,m}$ and $\boldsymbol{R}_{Rb,m}$ are used to denote the normalized transmit and receive spatial correlation.

The frequency correlation between subbands is modeled as follows. The channel between the BS and MS $m$ is characterized by a power delay profile [26]

$$S_m(\tau) = G_m \sum_{l=0}^{L_p-1} \frac{\sigma_{m,l}^2}{\Sigma_m} \delta(\tau - \tau_l), \tag{3}$$

with $L_p$ denoting the number of independent propagation paths, $\sigma_{m,l}^2$ and $\tau_l$ being, respectively, the power and delay of the $l$th propagation path, $\Sigma_m = \sum_{l=0}^{L_p-1} \sigma_{m,l}^2$, and

$$G_m = \varpi_m d_m^{-\varsigma_m} 10^{\chi_m/10} \Sigma_m \tag{4}$$

representing the average channel propagation gain between the BS and MS $m$, where $\varpi_m$ is a constant, $d_m$ is the distance between the BS and MS $m$, $\varsigma_m$ is the path-loss exponent, and $\chi_m \sim \mathcal{N}(0, \sigma_\chi^2)$ is a random variable modeling the shadow fading experienced by MS $m$. Hence, without taking into account the transmit and receive correlation between antennas and assuming that the channel coherence time is greater than $T_s$, the channel impulse response between transmit antenna $n_T$ and the receive antenna $n_R$ of MS $m$, over the whole frame period $t$, can be written as

$$\mathfrak{h}_m^{n_R,n_T}(t;\tau) = \sum_{l=0}^{L_p-1} \mathfrak{h}_{m,l}^{n_R,n_T}(t) \delta(\tau - \tau_l), \tag{5}$$

where $E\{|\mathfrak{h}_{m,l}^{n_R,n_T}(t)|^2\} = \varpi_m d_m^{-\varsigma_m} 10^{\chi_m/10}\sigma_{m,l}^2$. The corresponding frequency response, when evaluated over subband $b$ (with center frequency $f_b$), can be safely approximated by

$$\mathcal{H}_{b,m}^{n_R,n_T}(t) = \sum_{l=0}^{L_p-1} \mathfrak{h}_{m,l}^{n_R,n_T}(t) e^{-j2\pi f_b \tau_l}. \tag{6}$$

### *B. Modeling imperfect CSI*

In practical wireless communication systems, with time-varying fading radio channels, imperfect CSI at both the transmitter (CSIT) and the receiver (CSIR) may arise from a variety of sources such as, for example, channel estimation errors, quantization of the channel estimate in the feedback channel and/or outdated channel estimates with respect to the current channel response. By modeling such imperfections and taking them into account in the transceiver design, a robust high performance link can be achieved [27].

Typically, perfect CSIR can be safely assumed to be acquired via training pilot sequences that allow the estimation of the channel. CSIT, however, obtained by means of feedback from the receiver or from previous receive measurements assuming some kind of channel reciprocity, is far from perfect in most realistic situations. In this paper, a perfect CSIR will be assumed, and imperfect CSIT will consist of a nonzero channel mean and a channel covariance matrix, or equivalently a channel estimate and its estimation error covariance [27]. Thus, the relation between this mean-covariance CSIT and the MIMO channel in (1) will be modeled as

$$\boldsymbol{H}_{b,m}(t) = \zeta \overline{\boldsymbol{H}}_{b,m}(t) + \sqrt{1-\zeta^2}\, \boldsymbol{R}_{R_{b,m}}^{1/2} \boldsymbol{\psi}_{b,m}(t) \boldsymbol{R}_{T_{b,m}}^{1/2}, \tag{7}$$

where $\overline{\boldsymbol{H}}_{b,m}(t) = \boldsymbol{R}_{R_{b,m}}^{1/2} \overline{\boldsymbol{\mathcal{H}}}_{b,m}(t) \boldsymbol{R}_{T_{b,m}}^{1/2}$, $\zeta$ is a parameter that determines the quality of the channel estimate, and $\boldsymbol{\psi}_{b,m}(t) \sim \mathcal{CN}_{N_R^{(m)},N_T}\left(\boldsymbol{0}, G_m\left(\boldsymbol{I}_{N_R^{(m)}} \otimes \boldsymbol{I}_{N_T}\right)\right)$ is the unknown part in the CSIT.

### *C. Transmitter*

MIMO technology comprises a great variety of techniques that can be used to exploit the propagation paths between the $N_T$ transmit antennas at BS and the $N_R^{(m)}$ receive antennas at MS $m \in \mathcal{N}_m$ (see [28] for a review). When CSI is available at the transmitter and receiver sides, and multiplexing in the spatial domain is not used, the joint use of maximum ratio transmission (MRT) [29] at the transmitter and maximal ratio combining (MRC) at the receiver is known to provide optimum performance in the sense of maximizing the received SNR.

Let us assume that subband $b$ has been allocated to MS $m$ and that the BS uses an MRT scheme to exploit the spatial diversity provided by the MIMO channel. In this case, bits from the queue of MS $m$ are channel encoded and mapped onto a sequence of symbols drawn from the allocated normalized unit energy complex constellation (e.g., BPSK, QPSK, 16QAM, 64QAM). Furthermore, before the usual OFDM modulation steps on each transmit antenna (IFFT, cyclic prefix (CP) appending and up-conversion), the symbols are allocated power and processed in accordance with the MRT scheme. Denoting by $s_{b,m}^{(c,o)}(t)$ the symbol to be sent to MS $m$ over subcarrier $c \in \{1,\ldots,N_{sc}\}$ of subband $b$ and OFDM symbol $o \in \{1,\ldots,N_o\}$ during time slot $t$, then the corresponding $N_T \times 1$ transmitted vector can be written as

$$\boldsymbol{x}_{b,m}^{(c,o)} = \sqrt{\frac{p_{b,m}(t)}{N_{sc}}}\ \boldsymbol{v}_{b,m}(t) s_{b,m}^{(c,o)}(t), \tag{8}$$

where $p_{b,m}(t)$ is the power allocated to MS $m$ on subband $b$ during the time slot $t$ (within a given subband, power is uniformly allocated to subcarriers), and $\boldsymbol{v}_{b,m}(t) \in \mathbb{C}^{N_T \times 1}$ denotes the unit energy linear transmit filter used by the MRT transmission system, which, for the situation at hand, as stated in [27], is found to be the eigenvector associated with the maximum eigenvalue of the matrix $\zeta^2 \overline{\boldsymbol{H}}_{b,m}^H(t)\overline{\boldsymbol{H}}_{b,m}(t) + G_m\left(1-\zeta^2\right) \mathrm{tr}\left\{\boldsymbol{R}_{Rb,m}\right\} \boldsymbol{R}_{T\,b,m}^H$.

### *D. Receiver*

At the receiver side, as usual, ideal synchronization and sampling processes, and an OFDM cyclic prefix duration greater than the maximum delay spread of the channel impulse response are assumed. In this case, the received samples at the output of the $N_R^{(m)}$ fast Fourier transform (FFT) processing stages of MS $m$ are given by the $N_R^{(m)} \times 1$ complex valued vector

$$\boldsymbol{y}_{b,m}^{(c,o)}(t) = \boldsymbol{H}_{b,m}(t)\boldsymbol{x}_{b,m}^{(c,o)}(t) + \boldsymbol{\eta}_{b,m}^{(c,o)}(t) = \sqrt{\frac{p_{b,m}(t)}{N_{sc}}}\boldsymbol{\alpha}_{b,m}(t) s_{b,m}^{(c,o)}(t) + \boldsymbol{\eta}_{b,m}^{(c,o)}(t), \tag{9}$$

where $\boldsymbol{\alpha}_{b,m}(t) = \boldsymbol{H}_{b,m}(t)\boldsymbol{v}_{b,m}(t)$, and $\boldsymbol{\eta}_{b,m}^{(c,o)}(t) \sim \mathcal{CN}_{N_R^{(m)},1}\left(\boldsymbol{0}, \sigma_\eta^2 \boldsymbol{I}_{N_R^{(m)}}\right)$.

The complex valued vector $\boldsymbol{y}_{b,m}^{(c,o)}(t)$ is finally processed by the maximal ratio combiner producing a combined complex sample with an instantaneous SNR that can be expressed as

$$\gamma_{b,m}(t) = \frac{p_{b,m}(t)\delta_{b,m}(t)}{N_{sc}\sigma_\eta^2}, \tag{10}$$

where

$$\delta_{b,m}(t) = \boldsymbol{\alpha}_{b,m}^H(t)\boldsymbol{\alpha}_{b,m}(t). \tag{11}$$

## III. STATISTICALLY ROBUST DESIGN OF THE SRA UNIT

The SRA decisions at the BS will depend on the degree of knowledge of the CSI. For a given network configuration, characterized through the instantaneous SNR for each active MS in the system, the best performance would obviously be achieved with perfect CSI being available at both sides of the corresponding links. In practical communication systems, however, it is of paramount importance to properly model the CSI imperfections arising from channel estimation errors, outdated channel estimates and/or quantization of the channel estimate in the feedback channel, and to take them into account in the SRA unit in order to achieve a robust high performance design. In the following subsection, the CSI uncertainty effects on the instantaneous SNR are statistically characterized. This statistical model is used in the sequel first to propose a robust design for the PHY layer resource allocation schemes guarantying a certain average system performance, specified in terms of an outage probability bound, and second, to present a robust prediction approach of the DLC layer behaviour that allows a proper setting of the user priority coefficients used in the SRA algorithms.

### A. Statistical characterization of the SNR

When conditioned on the set of variables $\mathbf{\Phi}_{b,m}(t) = \left\{\zeta, \boldsymbol{R}_{Tb,m}, \boldsymbol{R}_{Rb,m}, \overline{\boldsymbol{\mathcal{H}}}_{b,m}(t), \boldsymbol{v}_{b,m}(t)\right\}$, the random vector $\boldsymbol{\alpha}_{b,m}(t)$ has a complex Gaussian distribution with mean $\zeta\overline{\boldsymbol{H}}_{b,m}(t)\boldsymbol{R}_{T_{b,m}}^{1/2}\boldsymbol{v}_{b,m}(t)$ and covariance matrix $\sigma_{Tb,m}(t)\,\boldsymbol{R}_{Rb,m}(t)$, where

$$\sigma_{Tb,m}(t) = G_m\left(1-\zeta^2\right)\mathrm{tr}\left\{\boldsymbol{R}_{T_{b,m}}^{1/2}\boldsymbol{v}_{b,m}(t)\boldsymbol{v}_{b,m}^H(t)\boldsymbol{R}_{T_{b,m}}^{H/2}\right\}. \tag{12}$$

Then, conditioned on $\mathbf{\Phi}_{b,m}(t)$, $\delta_{b,m}(t)$ is a noncentral complex quadratic form whose distribution can be denoted as [30, Definition III]

$$\delta_{b,m}(t)|\mathbf{\Phi}_{b,m}(t) \sim \mathcal{Q}_{1,N_R^{(m)}}\left(\boldsymbol{I}_{N_R^{(m)}}, \sigma_{Tb,m}(t),\ \boldsymbol{R}_{Rb,m}(t), \zeta\,\boldsymbol{v}_{b,m}^H(t)\overline{\boldsymbol{H}}_{b,m}^H(t)\right). \tag{13}$$

Thus, using [30, (5)] it can be shown that, conditioned on $\mathbf{\Phi}_{b,m}(t)$, the probability density function (pdf) of $\delta_{b,m}(t)$ is

$$p_{\delta_{b,m}(t)|\mathbf{\Phi}_{b,m}(t)}(x) = \frac{x^{N_R^{(m)}-1}\exp\left(-\frac{\omega x+\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2}{\sigma_{Tb,m}(t)}\right)}{\left[\sigma_{Tb,m}(t)\right]^{N_R^{(m)}}\left|\boldsymbol{R}_{Rb,m}\right|}\sum_{k=0}^{\infty}\frac{\tilde{P}_k\left(\boldsymbol{\mathcal{T}}_{b,m}(t), \boldsymbol{\mathcal{A}}_{b,m}(t), x/\sigma_{Tb,m}(t)\right)}{k!\Gamma\left(N_R^{(m)}+k\right)}, \tag{14}$$

where $\boldsymbol{u}_{b,m}(t) = \overline{\boldsymbol{\mathcal{H}}}_{b,m}(t)\boldsymbol{R}_{T_{b,m}}^{1/2}\boldsymbol{v}_{b,m}(t)$, $\omega \geq 0$ is an arbitrary constant,

$$\boldsymbol{\mathcal{T}}_{b,m}(t) = \frac{\zeta\, \boldsymbol{u}_{b,m}^H(t)}{\left[\sigma_{Tb,m}(t)\right]^{1/2}} \left(\boldsymbol{I}_{N_R^{(m)}} - \omega \boldsymbol{R}_{Rb,m}\right)^{-1/2}, \tag{15a}$$

$$\boldsymbol{\mathcal{A}}_{b,m}(t) = \boldsymbol{R}_{Rb,m}^{-1} - \omega \boldsymbol{I}_{N_R^{(m)}}, \tag{15b}$$

and $\tilde{P}_k(\cdot)$ is the generalized complex Hayakawa polynomial (corresponding to $k$) with two matrix arguments, defined in terms of complex zonal polynomials $\tilde{C}_k(\cdot)$ as [31], [32]

$$\tilde{P}_k(\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}) = E\left\{\tilde{C}_k\left(-\boldsymbol{\mathcal{B}}(\boldsymbol{V} - i\boldsymbol{\mathcal{T}})\boldsymbol{\mathcal{A}}(\boldsymbol{V} - i\boldsymbol{\mathcal{T}})^H\right)\right\}. \tag{16}$$

In our case, $\boldsymbol{\mathcal{T}} \in \mathbb{C}^{1\times N_R^{(m)}}$, $\boldsymbol{\mathcal{A}} \in \mathbb{C}^{N_R^{(m)}\times N_R^{(m)}}$ is Hermitian, and the expectation is taken with respect to $\boldsymbol{V} \in \mathbb{C}^{1\times N_R^{(m)}}$, with $\boldsymbol{V} \sim \mathcal{CN}_{1,N_R}\left(\boldsymbol{0}_{1\times N_R}, \boldsymbol{I}_{N_R^{(m)}}\right)$.

Using the definition of a complex zonal polynomial [33, (2)-(6)] we have that

$$\begin{aligned}\tilde{P}_k\left(\boldsymbol{\mathcal{T}}_{b,m}(t), \boldsymbol{\mathcal{A}}_{b,m}(t), x/\sigma_{Tb,m}(t)\right) &= \left(\frac{-x}{\sigma_{Tb,m}(t)}\right)^k \\ &\times E\left\{\left[(\boldsymbol{V} - i\boldsymbol{\mathcal{T}}_{b,m}(t))\boldsymbol{\mathcal{A}}_{b,m}(t)(\boldsymbol{V} - i\boldsymbol{\mathcal{T}}_{b,m}(t))^H\right]^k\right\}.\end{aligned} \tag{17}$$

Let us define $\boldsymbol{\mathcal{X}}_{b,m}(t) = (\boldsymbol{V} - i\boldsymbol{\mathcal{T}}_{b,m}(t))^H$, then

$$\boldsymbol{\mathcal{X}}_{b,m}(t) \sim \mathcal{CN}_{N_R^{(m)},1}\left(i\boldsymbol{\mathcal{T}}_{b,m}(t)^H, \boldsymbol{I}_{N_R^{(m)}}\right). \tag{18}$$

Let us also define the Hermitian quadratic form

$$\boldsymbol{\mathcal{Q}}\left(\boldsymbol{\mathcal{X}}_{b,m}(t)\right) = \boldsymbol{\mathcal{X}}_{b,m}^H(t)\boldsymbol{\mathcal{A}}_{b,m}(t)\boldsymbol{\mathcal{X}}_{b,m}(t). \tag{19}$$

Then, using [34, Chapter 4] and [35, Chapter 3], the moments of this quadratic form are obtained recursively as

$$\nu_{b,m}^{(k)}(t) = E\left\{\left[\boldsymbol{\mathcal{Q}}\left(\boldsymbol{\mathcal{X}}_{b,m}(t)\right)\right]^k\right\} = \sum_{r=0}^{k-1}\binom{k-1}{r} g_{b,m}^{(k-r)}(t)\nu_{b,m}^{(r)}(t), \tag{20}$$

with $\nu_{b,m}^{(0)}(t) = 1$, and

$$g_{b,m}^{(r)}(t) = r!\left[\frac{1}{2r}\text{tr}\left\{\hat{\boldsymbol{\mathcal{A}}}_{b,m}^r(t)\right\} + \hat{\boldsymbol{\mathcal{T}}}_{b,m}(t)\hat{\boldsymbol{\mathcal{A}}}_{b,m}^r(t)\hat{\boldsymbol{\mathcal{T}}}_{b,m}^T(t)\right], \tag{21}$$

where

$$\hat{\boldsymbol{\mathcal{T}}}_{b,m}(t) = \left[\mathbb{I}m\{\boldsymbol{\mathcal{T}}_{b,m}(t)\} \quad \mathbb{R}e\{\boldsymbol{\mathcal{T}}_{b,m}(t)\}\right], \tag{22}$$

and

$$\hat{\boldsymbol{\mathcal{A}}}_{b,m}(t) = \begin{bmatrix}\mathbb{R}e\{\boldsymbol{\mathcal{A}}_{b,m}(t)\} & -\mathbb{I}m\{\boldsymbol{\mathcal{A}}_{b,m}(t)\} \\ \mathbb{I}m\{\boldsymbol{\mathcal{A}}_{b,m}(t)\} & \mathbb{R}e\{\boldsymbol{\mathcal{A}}_{b,m}(t)\}\end{bmatrix}. \tag{23}$$

Thus, the pdf of $\delta_{b,m}(t)$, conditioned on $\mathbf{\Phi}_{b,m}(t)$, can be finally written as

$$p_{\delta_{b,m}(t)|\mathbf{\Phi}_{b,m}(t)}(x) = \frac{x^{N_R^{(m)}-1}\exp\left(-\frac{\omega x+\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2}{\sigma_{Tb,m}(t)}\right)}{\left[\sigma_{Tb,m}(t)\right]^{N_R^{(m)}}\left|\boldsymbol{R}_{Rb,m}\right|}\sum_{k=0}^{\infty}\frac{\nu_{b,m}^{(k)}(t)}{k!\Gamma\left(N_R^{(m)}+k\right)}\left(\frac{-x}{\sigma_{Tb,m}(t)}\right)^k. \quad (24)$$

For later use, let us also calculate the cdf of $\delta_{b,m}(t)$ conditioned on $\mathbf{\Phi}_{b,m}(t)$ as

$$F_{\delta_{b,m}(t)|\mathbf{\Phi}_{b,m}(t)}(x) = 1 - \frac{\exp\left(-\frac{\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2}{\sigma_{Tb,m}(t)}\right)}{\left|\boldsymbol{R}_{Rb,m}\right|}\sum_{k=0}^{\infty}\frac{(-1)^k\nu_{b,m}^{(k)}(t)\Gamma\left(N_R^{(m)}+k,\frac{\omega\, x}{\sigma_{Tb,m}(t)}\right)}{k!\Gamma\left(N_R^{(m)}+k\right)\omega^{N_R^{(m)}+k}}, \quad (25)$$

where $\Gamma(a,x)=\int_x^{\infty}t^{a-1}e^{-t}dt$ denotes the complementary incomplete Gamma function.

For the special case where the receive antennas are uncorrelated[2], i.e. $\boldsymbol{R}_{Rb,m}=\boldsymbol{I}_{N_R^{(m)}}$, we have that, using $\omega=1$,

$$\nu_{b,m}^{(k)}(t) = \left(\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2/\sigma_{Tb,m}(t)\right)^k. \quad (26)$$

Hence, rearranging terms, the following alternative equivalent closed-form of (24) can be obtained

$$\begin{aligned} p_{\delta_{b,m}(t)|\mathbf{\Phi}_{b,m}(t)}(x) =& \frac{\exp\left(-\frac{x+\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2}{\sigma_{Tb,m}(t)}\right)}{\sigma_{Tb,m}(t)}\left(\frac{x}{\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2}\right)^{\frac{N_R^{(m)}-1}{2}} \\ &\times I_{N_R^{(m)}-1}\left(\frac{2\sqrt{\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2 x}}{\sigma_{Tb,m}(t)}\right), \end{aligned} \quad (27)$$

where $I_n(\cdot)$ is the modified Bessel function of order $n$. Furthermore, using [36, (18)] we have that (25) reduces to

$$F_{\delta_{b,m}(t)|\mathbf{\Phi}_{b,m}(t)}(x) = 1 - Q_{N_R^{(m)}}\left(\sqrt{\frac{2\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2}{\sigma_{Tb,m}(t)}},\sqrt{\frac{2x}{\sigma_{Tb,m}(t)}}\right), \quad (28)$$

with $Q_n(a,b)$ denoting the $n$th-order (generalized) Marcum $Q$ function [37].

### *B. Statistically robust resource allocation*

*1) Power allocation:* Let $\boldsymbol{p}_b(t)=[p_{b,1}(t)\cdots p_{b,N_m}(t)]^T$ denote the vector of power allocation values for subband $b$ and time slot $t$. For a given set of constraints, the SRA algorithm will be in charge of determining the power allocation vector

$$\boldsymbol{p}(t) = \left[\left(\boldsymbol{p}_1(t)\right)^T\cdots\left(\boldsymbol{p}_{N_b}(t)\right)^T\right]^T \quad (29)$$

[2]This single-sided correlation model typically occurs when the MS is equipped with sufficiently spaced antennas, it is located in a rich scattering environment and communicates with an elevated BS with a limited number of surrounding scatterers [32].

optimizing a prescribed objective function. In addition to determining the power allocation values, the resource allocation algorithms should also allocate subbands and transmission rates. Nevertheless, as it will be shown next, the power allocation vector $\boldsymbol{p}(t)$ can also be used to represent the allocation of all these resources, thus simplifying the formulation of the optimization problem in Section IV [38].

*2) Subband allocation:* As usual, it is assumed that subband allocation (or RB) is exclusive, that is, only one MS is allowed to transmit on a given subband. Hence, the subband allocation constraints can be captured by constraining the power allocation vectors as

$$\boldsymbol{p}_b(t) \in \boldsymbol{\mathcal{P}}_b, \tag{30}$$

where $\boldsymbol{\mathcal{P}}_b \triangleq \left\{\boldsymbol{p}_b \in \mathbb{R}_+^{N_m} : p_{b,m}p_{b,m'} = 0,\ \forall\ m' \neq m\right\}$, with $\mathbb{R}_+$ denoting the set of all non-negative real numbers. Hence, the power allocation vector satisfies

$$\boldsymbol{p}(t) \in \boldsymbol{\mathcal{P}} = \boldsymbol{\mathcal{P}}_1 \times \cdots \times \boldsymbol{\mathcal{P}}_{N_b} \subset \mathbb{R}_+^{N_m N_b}, \tag{31}$$

where $\times$ denotes the Cartesian product (or product set).

*3) Robust rate allocation:* In the downlink of multi-rate systems based on AMC, a channel estimate is obtained at the receiver of each MS and it is then fed back to the BS so that the transmission scheme, comprising a modulation format and a channel code, can be adapted in accordance with the channel characteristics. If MS $m$ is allocated subband $b$ over time slot $t$, then the BS selects a modulation and coding scheme (MCS) that can be characterized by a transmission rate $\rho_{b,m}(t)$ (measured in bits per second). As each subband contains $N_{sc}$ subcarriers, the aggregated data rate allocated to MS $m$ on subband $b$ during time slot $t$ will be given by

$$r_{b,m}(t) = N_{sc}\rho_{b,m}(t). \tag{32}$$

*a) Discrete-rate AMC:* AMC strategies use a discrete set $\mathcal{N}_n = \{0, 1, \ldots, N_n\}$ of MCSs that can differ for different MSs. Each MCS is characterized by a particular transmission rate $\varrho_m^{(n)}$, with $\varrho_m^{(1)} < \ldots < \varrho_m^{N_n}$. The data rate $\varrho_m^{(0)} = 0$ corresponds to the no-transmission mode, that is, the mode selected when the channel is so bad that no bits can be transmitted to MS $m$ while guaranteeing the prescribed QoS constraints.

Transmission rate $\varrho_m^{(n)}$ can be related to block error rate (BLER) observed by MS $m$, denoted as $\epsilon_m$, and instantaneous SNR $\gamma_{b,m}(t)$ as [39, Chapter 9] (see also [24])

$$\epsilon_m(\gamma_{b,m}(t), \varrho_m^{(n)}) = \begin{cases} 1, & \gamma_{b,m}(t) < \gamma_m^{(n)} \\ \kappa_1^{(n)} \exp\left(-\frac{\kappa_2^{(n)}\gamma_{b,m}(t)}{2^{T_o\varrho_m^{(n)}}-1}\right), & \text{otherwise,} \end{cases} \tag{33}$$

where $\kappa_1^{(n)}$ and $\kappa_2^{(n)}$ are modulation- and code-specific constants that can be accurately approximated by exponential curve fitting. This expression is general enough to obtain the BLER performance of any transmission system for which the joint effects of transmission filters, channel coefficients and reception filters can be represented through an instantaneous SNR $\gamma_{b,m}(t)$, which for the special case of MRT/MRC scheme is given by (10).

In a system with ideal CSIT, given $p_{b,m}(t)$, $\delta_{b,m}(t)$ and $\sigma_\eta^2$, the transmitter can use (10) to find the instantaneous SNR $\gamma_{b,m}(t)$ and then, assuming a maximum allowable BLER $\check{\epsilon}_m$, employ (33) to select the most adequate MCS scheme as the one with transmission rate

$$\rho_{b,m}(t) = \max\left\{\varrho_m^{(n)} \;:\; \epsilon\left(\gamma_{b,m}(t), \varrho_m^{(n)}\right) \leq \check{\epsilon}_m\right\}. \tag{34}$$

In this case, the transmission rate $\rho_{b,m}(t)$ can be expressed using the staircase function

$$\rho_{b,m}(t) = \begin{cases} \varrho_m^{(0)}, & 0 \leq \gamma_{b,m}(t) < \Gamma_m^{(1)} \\ \varrho_m^{(1)}, & \Gamma_m^{(1)} \leq \gamma_{b,m}(t) < \Gamma_m^{(2)} \\ & \vdots \\ \varrho_m^{(N_n)}, & \Gamma_m^{(N_n)} \leq \gamma_{b,m}(t) < \infty \end{cases} \tag{35}$$

where $\left\{\Gamma_m^{(n)}\right\}_{n=1}^{N_n-1}$, with $\Gamma_m^{(n)} \leq \Gamma_m^{(n+1)}$, are the instantaneous SNR boundaries defining the MCS intervals, which can be obtained from (33) as

$$\Gamma_m^{(n)} = \frac{2^{T_o \varrho_m^{(n)}} - 1}{\kappa_2^{(n)}} \ln \frac{\kappa_1^{(n)}}{\check{\epsilon}_m}. \tag{36}$$

In a system with imperfect CSIT, however, the transmitter does not know the value of $\delta_{b,m}(t)$. In fact, it has only access to the set of variables $\mathbf{\Phi}_{b,m}(t)$. In this case, a robust transmitter design can guarantee a BLER outage probability $\check{\pi}_m$ associated to a maximum allowable BLER $\check{\epsilon}_m$. That is, given $p_{b,m}(t)$, $\sigma_\eta^2$ and $\mathbf{\Phi}_{b,m}(t)$, the transmitter can select the optimal MCS scheme as the one with transmission rate

$$\begin{aligned} \rho_{b,m}(t) &= \max\left\{\varrho_m^{(n)} : \Pr\left\{\epsilon\left(\gamma_{b,m}(t), \varrho_m^{(n)}\right) \geq \check{\epsilon}_m\right\} \leq \check{\pi}_m\right\} \\ &= \max\left\{\varrho_m^{(n)} : F_{\delta_{b,m}(t)|\mathbf{\Phi}_{b,m}(t)}\left(\frac{N_{sc}\sigma_\eta^2 \Gamma_m^{(n)}}{p_{b,m}(t)}\right) \leq \check{\pi}_m\right\}. \end{aligned} \tag{37}$$

Thus, as for the ideal CSIT case, the transmission rate $\rho_{b,m}(t)$ can be expressed using the staircase function in (35) by substituting $\gamma_{b,m}(t)$ with

$$\hat{\gamma}_{b,m}(t) = \frac{p_{b,m}(t)}{N_{sc}\sigma_\eta^2} F^{-1}_{\delta_{b,m}(t)|\mathbf{\Phi}_{b,m}(t)}\left(\check{\pi}_m\right). \tag{38}$$

*b) Continuous-rate AMC:* A useful abstraction when exploring rate limits is to assume that each user's set of MCSs is infinite. In this case, for each value of $\gamma \geq 0$ there is an MCS characterized with a transmission rate

$$\varrho_m(\gamma) = \frac{1}{T_o} \log_2 \left(1 + \frac{\gamma}{\Lambda_m}\right), \tag{39}$$

and a BLER given by

$$\epsilon\left(\gamma_{b,m}(t), \varrho_m(\gamma)\right) = \begin{cases} 1, & \gamma_{b,m}(t) < \gamma \\ 0, & \text{otherwise}, \end{cases} \tag{40}$$

where $\Lambda_m \geq 1$ represents the coding gap due to the utilization of a practical (rather than ideal) coding scheme. With $\Lambda_m = 1$ this expression results in the Shannon's capacity limit and allows the comparison of practical AMC-based schemes against fundamental capacity-achieving benchmarks.

In a system with ideal CSIT, given the instantaneous SNR $\gamma_{b,m}(t)$, the transmitter can use (40) to select the most adequate MCS scheme as the one with transmission rate

$$\rho_{b,m}(t) = \max\left\{\varrho_m(\gamma) \,:\, \epsilon\left(\gamma_{b,m}(t), \varrho_m(\gamma)\right) \leq \check{\epsilon}_m\right\} = \frac{1}{T_o} \log_2 \left(1 + \frac{p_{b,m}(t)\delta_{b,m}(t)}{N_{sc}\sigma_\eta^2 \Lambda_m}\right). \tag{41}$$

In a system with imperfect CSIT, the maximum allowable transmission rate fulfilling the prescribed BLER constraints can be obtained as

$$\begin{aligned} \rho_{b,m}(t) &= \max\left\{\varrho_m(\gamma) : \Pr\left\{\epsilon\left(\gamma_{b,m}(t), \varrho_m(\gamma)\right) \geq \check{\epsilon}_m\right\} \leq \check{\pi}_m\right\} \\ &= \max\left\{\varrho_m(\gamma) : F_{\delta_{b,m}(t)|\mathbf{\Phi}_{b,m}(t)}\left(\frac{N_{sc}\sigma_\eta^2\gamma}{p_{b,m}(t)}\right) \leq \check{\pi}_m\right\} \\ &= \frac{1}{T_o} \log_2 \left(1 + \frac{p_{b,m}(t) F^{-1}_{\delta_{b,m}(t)|\mathbf{\Phi}_{b,m}(t)}\left(\check{\pi}_m\right)}{N_{sc}\sigma_\eta^2 \Lambda_m}\right). \end{aligned} \tag{42}$$

### *C. Statistically robust scheduling predictions*

As shown in [11], most of the SRA schemes that have been proposed in the literature can be interpreted as decision making algorithms that, at the beginning of time slot $t$ estimate or predict the future behavior of QoS quantitative performance measures such as the throughput, average delay, queue length and/or head-of-line delay, and decide which users will be granted a transmission opportunity and the amount of resources that they will be allocated.

At the beginning of time slot $t$, MS $m$ is assumed to have $Q_m(t)$ bits in the queue. If there are $A_m(t)$ bits arriving during time slot $t$, the queue length at the end of this time slot, assuming queues of infinite capacity[3], can then be expressed as [8, Section IV.A]

$$Q_m(t+1) = Q_m(t) + A_m(t) - R_m(t)\, N_o T_o, \tag{43}$$

where

$$R_m(t) = \min\left\{ \sum_{b=1}^{N_b} \vartheta_{b,m}(t) r_{b,m}(t),\ \sum_{b=1}^{N_b} \vartheta_{b,m}(t) \frac{q_{b,m}(t)}{N_o T_o} \right\}, \tag{44}$$

with

$$\vartheta_{b,m}(t) = \begin{cases} 1 & \text{successful transmission of RB} \\ 0 & \text{otherwise} \end{cases} \tag{45}$$

and $\sum_{b=1}^{N_b} q_{b,m}(t) = Q_m(t)$, that is, $q_{b,m}(t)$ denotes the quantity of information that is transmitted by MS $m$ on RB $b$ when $Q_m(t) \leq N_{sc} N_o T_o \sum_{b=1}^{N_b} \rho_{b,m}(t)$.

*1) Predicting the queue length:* As $A_m(t)$ and $\boldsymbol{\vartheta}_m(t) = \{\vartheta_{b,m}(t)\}_{b=1}^{N_b}$ are unknown at the beginning of time slot $t$, and assuming that the DLC layer only knows the average arrival data rate $\lambda_m$, then a prediction of the queue length at the end of this time slot can be obtained from (43) as

$$\hat{Q}_m(t+1) = E_{A_m, \boldsymbol{\vartheta}_m}\{Q_m(t+1)\} = Q_m(t) + \lambda_m T_s - \overline{R}_m(t)\, N_o T_o, \tag{46}$$

where $E_{\boldsymbol{x}}\{\cdot\}$ denotes the statistical expectation operator with respect to the random variables in vector $\boldsymbol{x}$ and, thus,

$$\overline{R}_m(t) = \min\left\{ \sum_{b=1}^{N_b} \overline{\vartheta}_{b,m}(t) r_{b,m}(t),\ \sum_{b=1}^{N_b} \overline{\vartheta}_{b,m}(t) \frac{q_{b,m}(t)}{N_o T_o} \right\}, \tag{47}$$

with

$$\overline{\vartheta}_{b,m}(t) = 1 - \epsilon_m(\gamma_{b,m}(t), \rho_{b,m}(t)), \tag{48}$$

for the ideal CSIT case, and

$$\overline{\vartheta}_{b,m}(t) = 1 - E_{\delta_{b,m}(t)|\boldsymbol{\Phi}_{b,m}(t)}\left\{\epsilon_m(\gamma_{b,m}(t), \rho_{b,m}(t))\right\}, \tag{49}$$

for the partial CSIT case.

[3]With a slight increase in notational complexity, this framework can be easily extended to the case with finite capacity queues by taking into account the overflown information bits when the queues are found full.

In a continuous allocation system (CRA) using an MCS with $\rho_{b,m}(t) = \varrho_m(\gamma)$ the BLER can be expressed as in (40) and, thus, $\overline{\vartheta}_{b,m}(t) = 1$ for the ideal CSIT case, and

$$\begin{aligned}\overline{\vartheta}_{b,m}(t) &= 1 - E_{\delta_{b,m}(t)|\boldsymbol{\Phi}_{b,m}(t)}\left\{\epsilon_m(\gamma_{b,m}(t), \rho_m(\gamma))\right\} \\ &= 1 - F_{\delta_{b,m}(t)|\boldsymbol{\Phi}_{b,m}(t)}\left(\frac{N_{sc}\sigma_\eta^2\gamma}{p_{b,m}(t)}\right) = 1 - \breve{\pi}_m,\end{aligned} \tag{50}$$

for the partial CSIT case.

On the other hand, in a discrete allocation system (DRA) using MCS $n$, the BLER can be expressed as in (33), hence

$$\overline{\vartheta}_{b,m}(t) = \overline{\vartheta}_{b,m}^{(n)}(t) = 1 - \kappa_1^{(n)} \exp\left(-\frac{\kappa_2^{(n)}\gamma_{b,m}(t)}{2^{T_o\varrho_m^{(n)}} - 1}\right), \tag{51}$$

for the ideal CSIT case, and

$$\begin{aligned}\overline{\vartheta}_{b,m}(t) &= \overline{\vartheta}_{b,m}^{(n)}(t) = 1 - E_{\delta_{b,m}(t)|\boldsymbol{\Phi}_{b,m}(t)}\left\{\epsilon_m(\gamma_{b,m}(t), \varrho_m^{(n)})\right\} \\ &= 1 - F_{\delta_{b,m}(t)|\boldsymbol{\Phi}_{b,m}(t)}\left(\frac{N_{sc}\sigma_\eta^2\gamma_m^{(n)}}{p_{b,m}(t)}\right) - \frac{\kappa_1^{(n)}\exp\left(-\frac{\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2}{\sigma_{Tb,m}(t)}\right)}{\left[B_{b,m}^{(n)}(t)\right]^{N_R^{(m)}}\left|\boldsymbol{R}_{Rb,m}\right|} \\ &\times \sum_{k=0}^{\infty}\frac{(-1)^k\nu_{b,m}^{(k)}(t)\Gamma\left(N_R^{(m)} + k, \frac{N_{sc}\sigma_\eta^2\gamma_m^{(k)}B_{b,m}^{(n)}(t)}{p_{b,m}(t)\sigma_{Tb,m}(t)}\right)}{k!\Gamma\left(N_R^{(m)} + k\right)\left[B_{b,m}^{(n)}(t)\right]^k}\end{aligned} \tag{52}$$

for the partial CSIT case, where

$$B_{b,m}^{(n)}(t) = \frac{\kappa_2^{(n)}\sigma_{Tb,m}(t)p_{b,m}(t)}{N_{sc}\sigma_\eta^2\left(2^{T_o\varrho_m^{(n)}} - 1\right)} + \omega. \tag{53}$$

For uncorrelated receive antennas, i.e. $\boldsymbol{R}_{Rb,m}(t) = \boldsymbol{I}_{N_R^{(m)}}$, (52) simplifies to (with $\omega = 1$)

$$\begin{aligned}\overline{\vartheta}_{b,m}(t) &= \overline{\vartheta}_{b,m}^{(k)}(t) = 1 - E_{\delta_{b,m}(t)|\boldsymbol{\Phi}_{b,m}(t)}\left\{\epsilon_m(\gamma_{b,m}(t), \rho_m^{(k)})\right\} \\ &= 1 - F_{\delta_{b,m}(t)|\boldsymbol{\Phi}_{b,m}(t)}\left(\frac{N_{sc}\sigma_\eta^2\gamma_m^{(k)}}{p_{b,m}(t)}\right) \\ &- \frac{\kappa_1^{(k)}}{\left[B_{b,m}^{(n)}(t)\right]^{N_R^{(m)}}}\exp\left(-\frac{\kappa_2^{(k)}\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2 p_{b,m}(t)}{N_{sc}\sigma_\eta^2\left(2^{T_o\varrho_m^{(k)}} - 1\right)B_{b,m}^{(n)}(t)}\right) \\ &\times Q_{N_R^{(m)}}\left(\sqrt{\frac{2\zeta^2\left\|\boldsymbol{u}_{b,m}(t)\right\|^2}{\sigma_{Tb,m}(t)B_{b,m}^{(n)}(t)}}, \sqrt{\frac{2B_{b,m}^{(n)}(t)N_{sc}\sigma_\eta^2\gamma_m^{(k)}}{p_{b,m}(t)\sigma_{Tb,m}(t)}}\right).\end{aligned} \tag{54}$$

*2) Predicting the head-of-line delay:* The HOL delay of user $m$ at the beginning of time slot $t$ (or equivalently, the end of time slot $(t-1)$) can be written as $W_{\text{HOL},m}(t) = t\,T_s - \tau_m^{(A)}(t)$, where $\tau_m^{(A)}(t)$ denotes the arrival time of the HOL packet to the queue of user $m$. Hence, the predicted HOL delay at the end of time slot $t$ can be readily obtained as

$$\begin{aligned}\hat{W}_{\text{HOL},m}(t+1) &= (t+1)T_s - \hat{\tau}_m^{(A)}(t+1) \\ &= (t+1)T_s - \left(\tau_m^{(A)}(t) + \frac{\overline{R}_m(t)N_oT_o}{\lambda_m}\right) \\ &= W_{\text{HOL},m}(t) + T_s - \frac{\overline{R}_m(t)N_oT_o}{\lambda_m}.\end{aligned} \tag{55}$$

## IV. Unified optimization framework

The main objective of cross-layer SRA algorithms over a wireless network is the establishment of effective policies able to optimize metrics related to spectral/energy efficiency and fairness, while satisfying prescribed QoS constraints. The issues of efficient and fair allocation of resources have been intensively investigated in the context of economics, where utility functions have been used to quantify the benefit obtained from the usage of a pool of resources. In a similar way, utility theory can be used in wireless communication networks to evaluate the degree up to which a given network configuration can satisfy users' QoS requirements [4], [5]. Invoking the procedure presented in [11] (and references therein), the cross-layer SRA problem at hand can be formally posed as [4]

$$\begin{aligned}&\max_{\boldsymbol{p}\in\boldsymbol{\mathcal{P}}} \sum_{m=1}^{N_m} w_m \overline{r}_m \\ &\text{subject to } \sum_{m=1}^{N_m}\sum_{b=1}^{N_b} p_{b,m} \le P_T,\end{aligned} \tag{56}$$

where $\overline{r}_m = \sum_{b=1}^{N_b} \overline{\vartheta}_{b,m} r_{b,m}$ is the effective date rate allocated to user $m$, and the $w_m$'s are the weighing coefficients used to implement different schedulers based on the predicted parameters. Details on how to set the weighing coefficients to follow prescribed scheduling rules can be found in [11]. For completeness, weights for some popular schedulers are given next.

[4]Note that since optimization is performed on a RB-by-RB basis, from this point onwards the time dependence (i.e., $(t)$) of all the variables will be dropped.

- *Proportional fair* (PF) rule [40] is based on a channel-aware scheduling rule aiming at maximizing the logarithmic-sum-throughput of the system. It is often implemented using weighs defined by

$$w_m = 1/\tilde{\bar{r}}_m, \ \forall\, m, \tag{57}$$

where $\tilde{\bar{r}}_m$ is the average effective data rate actually allocated to user $m$, which is calculated using a moving average over a relatively long sliding window [41].

- *Modified largest weighted delay first (MLWDF)* [42] is based on a channel- and queue-aware scheduling rule. At each time slot $t$, the MLWDF scheduler aims at choosing the best combination of queueing delay and potential transmission rate by setting the weights to [42],

$$w_m = \phi_m W_{\text{HOL},m}/\tilde{\bar{r}}_m, \ \forall\, m, \tag{58}$$

where $\phi_m$ are arbitrary positive constants that can be used to set different priority levels between traffic flows. In order to guarantee that users with absolute delay requirement $\check{D}_m$ and maximum outage delay probability requirement $\check{\xi}_m$ will be satisfied, the authors of [42] propose to *properly* set the values of $\phi_m$ as

$$\phi_m = -\frac{\log(\check{\xi}_m)}{\check{D}_m}, \tag{59}$$

providing in this way QoS differentiation among flows.

- *Exponential (EXP)* rule [43] is also based on a channel- and queue-aware throughput optimal scheduling rule that considers the waiting time in the queues, the instantaneous potential transmission rates and the maximum tolerable delay requirements. The weights in this case can be shown to be defined by

$$w_m = \frac{\phi_m}{\tilde{\bar{r}}_m} \exp\left(\frac{\phi_m W_{\text{HOL},m} - \overline{\phi W}}{1 + \sqrt{\overline{\phi W}}}\right)$$

for all $m$, with

$$\overline{\phi W} = \frac{1}{N_m} \sum_{m=1}^{N_m} \phi_m W_{\text{HOL},m}. \tag{60}$$

The optimization problem formulated in (56) is general enough to account for different power and rate allocation strategies: uniform power allocation (UPA), adaptive power allocation (APA), continuous rate allocation (CRA) and discrete rate allocation (DRA), which are next treated in detail.

### A. Uniform Power Allocation

Let us assume that the BS transmit power $P_T$ is uniformly allocated to all subbands. In this case, if subband $b$ is allocated to user $m_b^*$, then the subband exclusive allocation constraint (i.e., $\boldsymbol{p} \in \boldsymbol{\mathcal{P}}$) forces that

$$p_{b,m} = \begin{cases} P_T/N_b, & m = m_b^* \\ 0, & m \neq m_b^*, \end{cases} \tag{61}$$

for all $b$. Thus, using (32) in (56) it is straightforward to show that subband $b$ must be allocated to MS $m_b^*$ satisfying

$$m_b^* = \arg \max_{m \in \mathcal{N}_m} \left\{ w_m \overline{\vartheta}_{b,m} \rho_{b,m} \right\}, \ \forall b, \tag{62}$$

with $\rho_{b,m}$ obtained as in (35) (see also (38)), for the DRA case, or (41) and (42), for the CRA case.

### B. Adaptive Power Allocation

The objective function in (56) is concave, but $\boldsymbol{\mathcal{P}}$ is a highly non-convex discrete constraint space. Fortunately, problem (56) is separable across subbands and, as stated in [38], [44], it can be approached by using Lagrange duality principles. With $\mu$ denoting the Lagrange multiplier associated with the power constraint, the Lagrangian of (56) can be expressed as

$$\mathcal{L}\left(\boldsymbol{p}, \mu\right) = \sum_{m=1}^{N_m} w_m \sum_{b=1}^{N_b} \overline{\vartheta}_{b,m} r_{b,m} + \mu \left( P_T - \sum_{m=1}^{N_m} \sum_{b=1}^{N_b} p_{b,m} \right), \tag{63}$$

and the dual problem can then be written as [45]

$$g(\boldsymbol{p}, \mu) = \min_{\mu \geq 0} \left\{ \max_{\boldsymbol{p} \in \boldsymbol{\mathcal{P}}} \mathcal{L}\left(\boldsymbol{p}, \mu\right) \right\} = \min_{\mu \geq 0} \left\{ \max_{\boldsymbol{p} \in \boldsymbol{\mathcal{P}}} \left[ \sum_{m=1}^{N_m} \sum_{b=1}^{N_b} \left( w_m \overline{\vartheta}_{b,m} r_{b,m} - \mu p_{b,m} \right) \right] + \mu P_T \right\}. \tag{64}$$

Now, using the subband exclusive allocation constraint and the separability of power variables across subbands, the dual problem can be simplified as

$$g(\boldsymbol{p}, \mu) = \min_{\mu \geq 0} \left\{ \sum_{b=1}^{N_b} \max_{\substack{m \in \mathcal{N}_m \\ p_{b,m} \geq 0}} \left\{ w_m \overline{\vartheta}_{b,m} r_{b,m} - \mu p_{b,m} \right\} + \mu P_T \right\}. \tag{65}$$

The solution to the simplified dual problem is given by optimizing (65) over all $(\boldsymbol{p}, \mu) \succeq 0$. This optimization can be done iteratively and coordinate-wise, starting with the $\boldsymbol{p}$ variables and continuing with $\mu$.

*1) Optimizing the dual function over $\boldsymbol{p}$:*

*a) Continuous rate allocation (CRA):* For a given value of $\mu$, in case of using $\rho_{b,m}$ as defined in either (41), for the ideal CSIT case, or (42), for the imperfect CSIT case, the innermost maximization in (65) provides a multilevel water-filling closed-form expression for the optimal power allocation given by

$$p_{b,m}^{*}=\left[\frac{N_{sc}w_m\overline{\vartheta}_{b,m}}{\mu T_o\ln 2}-\frac{N_{sc}\Lambda_m\sigma_\eta^2}{\varphi_{b,m}}\right]^{+}, \tag{66}$$

where $[x]^{+}\triangleq\max\{0,x\}$ and[5]

$$\varphi_{b,m}=\begin{cases}\delta_{b,m} & \text{for perfect CSIT}\\ F^{-1}_{\delta_{b,m}|\boldsymbol{\Phi}_{b,m}}\left(\breve{\pi}_m\right) & \text{for imperfect CSIT.}\end{cases} \tag{67}$$

Now, using (66) in (65) yields

$$g(\mu)=\min_{\mu\geq 0}\left\{\sum_{b=1}^{N_b}\max_{m\in\mathcal{N}_m}\left\{w_m\overline{\vartheta}_{b,m}r_{b,m}^{*}-\mu p_{b,m}^{*}\right\}+\mu P_T\right\}, \tag{68}$$

where

$$r_{b,m}^{*}=\frac{N_{sc}}{T_o}\log_2\left(1+\frac{p_{b,m}^{*}\varphi_{b,m}}{\sigma_\eta^2 N_{sc}\Lambda_m}\right). \tag{69}$$

Hence, for a fixed dual variable $\mu$, the subband $b$ will be allocated to MS $m_b^{*}$ satisfying

$$m_b^{*}=\arg\max_{m\in\mathcal{N}_m}\left\{w_m\overline{\vartheta}_{b,m}r_{b,m}^{*}-\mu p_{b,m}^{*}\right\},\forall b. \tag{70}$$

*b) Discrete rate allocation (DRA):* In this case, the innermost maximization in (65) has to deal with a non-derivable discontinuous objective function. However, an approach derived from that proposed by Wong and Evans in [38, Chapter 3] can be applied to arrive at either the optimal or an approximate suboptimal solution. Using (35) the set of non-negative real numbers (i.e., $\mathbb{R}^{+}$) can be subdivided, for each MS $m$ and subband $b$, into the $N_n$ segments

$$\mathcal{R}_{b,m,n}^{+}=\left[\frac{N_{sc}\sigma_\eta^2\Gamma_m^{(n)}}{\varphi_{b,m}},\frac{N_{sc}\sigma_\eta^2\Gamma_m^{(n+1)}}{\varphi_{b,m}}\right),\ n\in\mathcal{N}_n. \tag{71}$$

Furthermore, it can be shown that, over each of these segments, the optimization subproblem is a standard concave maximization over a convex set with a unique solution given by

$${p_{b,m}^{(n)}}^{*}=\arg\max_{p_{b,m}\in\mathcal{R}_{b,m,n}^{+}}\left\{w_m N_{sc}\overline{\vartheta}_{b,m}^{(n)}\varrho_m^{(n)}-\mu p_{b,m}\right\}, \tag{72}$$

[5]Subsequent expressions involving (67) should be understood to be valid for any type of CSIT just by choosing the corresponding entry for $\varphi_{b,m}$.

which can be found by a standard line-search algorithm. As a consequence, there only exist $N_n$ candidate power allocations

$$p_{b,m}^{*} \in \left\{ {p_{b,m}^{(0)}}^{*}, {p_{b,m}^{(1)}}^{*}, \ldots, {p_{b,m}^{(N_n-1)}}^{*} \right\} \tag{73}$$

from which the one maximizing $w_m N_{sc} {\overline{\vartheta}_{b,m}^{(n)}}^{*} \varrho_m^{(n)} - \mu {p_{b,m}^{(n)}}^{*}$ must be selected, that is,

$$p_{b,m}^{*} = {p_{b,m}^{(n_{b,m}^{*})}}^{*}, \tag{74}$$

where

$$n_{b,m}^{*} = \arg\max_{n \in \mathcal{N}_n} \left\{ w_m N_{sc} {\overline{\vartheta}_{b,m}^{(n)}}^{*} \varrho_m^{(n)} - \mu {p_{b,m}^{(n)}}^{*} \right\}. \tag{75}$$

Furthermore, as in the CRA case, given $\mu$, $n_{b,m}^{*}$ and $p_{b,m}^{*}$, the subband $b$ must be allocated to MS $m_b^{*}$ satisfying

$$m_b^{*} = \arg\max_{m \in \mathcal{N}_m} \left\{ w_m N_{sc} \overline{\vartheta}_{b,m}^{(n_{b,m}^{*})} \varrho_m^{(n_{b,m}^{*})} - \mu p_{b,m}^{*} \right\}, \forall b. \tag{76}$$

An approximate, less complex, suboptimal solution can be obtained if instead of using (51) or (52) to obtain the exact expression for $\overline{\vartheta}_{b,m}^{(n)}$, we use the approximations[6]

$$\tilde{\vartheta}_{b,m}^{(n)} = \begin{cases} 1 & \text{for ideal CSIT} \\ 1 - \check{\pi}_m & \text{for partial CSIT,} \end{cases} \tag{77}$$

which are not dependent on the value of $p_{m,b}$. In these cases, if a power allocation $p_{b,m} \in \mathcal{R}_{b,m,n}^{+}$ is used, then (72) can be approximated as

$${p_{b,m}^{(n)}}^{*} \simeq N_{sc} \sigma_\eta^2 \Gamma_m^{(n)} / \varphi_{b,m}. \tag{78}$$

*2) Optimizing the dual function over $\mu$:* Once known the optimal vector $\boldsymbol{p}^{*}$ for a given $\mu$, the dual optimization problem (65) reduces to

$$g(\mu) = \min_{\mu \geq 0} \left\{ \sum_{b=1}^{N_b} \left( w_m \overline{\vartheta}_{b,m}^{*} N_{sc} \rho_{m_b^{*},b}^{*} - \mu p_{m_b^{*},b}^{*} \right) + \mu P_T \right\}. \tag{79}$$

Using standard properties of dual optimization problems [38], [44], it can be shown that this problem is convex with respect to $\mu$, and thus, derivative-free line search methods like, for example, Golden-section or Fibonacci, can be used to determine $\mu^{*}$. Once $\mu^{*}$ has been found, it can be used to obtain optimal power, subband and rate allocation for each of the data flows in the system.

[6]These are excellent approximations when using AMC schemes based on powerful Turbo or low-density parity-check (LDPC) codes characterized with very steep slopes in the waterfall region of the BLER curves. For instance, Turbo codes showing these characteristics have been standardized by 3GPP for use in the evolved-UTRAN (E-UTRAN) radio access of Long Term Evolution (LTE) and LTE-Advanced (LTE-A) systems [21], [22].

## V. NUMERICAL RESULTS

This section, using the unified cross-layer framework introduced in previous sections, presents results that serve to illustrate how this can be used to evaluate different SRA policies under different CSIT conditions. To this end, a unicellular downlink MIMO-OFDMA wireless network with a cell radius of 500 m has been simulated with a BS serving a set of MSs uniformly distributed over the whole coverage area. System parameters have been defined based on current LTE specifications [46]. In particular, the entire system bandwidth is $B = 10$ MHz, and is divided into $N_b = 50$ orthogonal subbands, each with a bandwidth $B_b = 16.67$ kHz and consisting of $N_{sc} = 12$ adjacent subcarriers. Transmission between the BS and active MSs is organized in time slots of duration $T_s = 0.46$ ms with each of these slots consisting of $N_o = 7$ OFDM symbols of duration $T_o = 61,13$ $\mu$s (excluding the cyclic prefix[7]). Thus, the basic resource unit is formed by 12 adjacent subcarriers (one subband) over 7 OFDM symbols (one time slot), usually known as a resource block. However, and following the LTE scheduling procedure, the SRA process takes place over two consecutive RBs (RB pair). A wireless channel including the path-losses, shadowing effects and frequency-, time- and space-selective fading experienced by the transmitted signal on its way from the BS to the MSs, has been generated conforming to the Extended Typical Urban (ETU) channel model defined within LTE [47] with a shadow fading standard deviation of 6 dB. For clarity, three different scenarios have been considered, namely, homogeneous and heterogeneous traffic conditions are covered in Scenarios 1 and 2, respectively, while the third scenario is devoted to the study of correlation effects. For all scenarios, and without loss of generality, the BS is assumed to have two transmit antennas ($N_T = 2$) whereas MSs are also assumed to have two receive antennas ($N_R = 2$). Moreover, correlation matrices are assumed to be common to all MSs and across the different subbands ($\boldsymbol{R}_{Rb,m} = \boldsymbol{R}_R, \boldsymbol{R}_{Tb,m} = \boldsymbol{R}_T \ \forall b, m$) and of the form

$$\boldsymbol{R}_R = \begin{pmatrix} 1 & \rho_{rx} \\ \rho_{rx} & 1 \end{pmatrix} \quad \boldsymbol{R}_T = \begin{pmatrix} 1 & \rho_{tx} \\ \rho_{tx} & 1 \end{pmatrix}. \tag{80}$$

In Scenarios 1 and 2 the correlation of the Tx and Rx antenna arrays is defined by setting $\rho_{tx} = \rho_{rx} = 0.1$. In Scenario 3, the correlation factor is the parameter under study.

[7]Note that in LTE, the first OFDM symbol within a slot has a CP duration slightly longer than the rest. In this work, and for the sake of implementation simplicity, all OFDM symbols forming a slot are assumed to have a CP with a duration of 4.7 $\mu$s.

The outage and maximum BLER have been set to $\breve{\pi}_m = \breve{\epsilon}_m = 0.01$ for all scenarios. For the simulation results in which discrete rate adaptation (DRA) is used, MCSs defined within LTE are employed whose transmission rates $\varrho_m^{(n)}$ are specified in Table 7.2.3-1 in [48] and lie in the range $(0.15, 5.55)$ bits/s/Hz. The parameters $\kappa_1^{(n)}$, $\kappa_2^{(n)}$ and $\gamma_m^{(n)}$ in (33) have been extracted from [19] and the approximation given by (77) has been used because, as it is shown in [19], it entails a negligible performance loss with respect to the use of the more complex (52)-(54). To illustrate the merits of the proposed robust design (based on an equivalent channel gain defined by $F^{-1}_{\delta_{b,m}|\boldsymbol{\Phi}_{b,m}}(\breve{\pi}_m)$), we will compare it against a *naive* cross-layer design in which SRA is based on the equivalent channel gain defined by $\bar{\delta}_{b,m}(t) = \|\overline{\boldsymbol{H}}_{b,m}(t)\bar{\boldsymbol{v}}_{b,m}(t)\|^2$ where $\bar{\boldsymbol{v}}_{b,m}(t)$ is the eigenvector associated to the largest eigenvalue of matrix $\overline{\boldsymbol{H}}^H_{b,m}(t)\overline{\boldsymbol{H}}_{b,m}(t)$. The naive character of this strategy stems from the fact that it neglects the imperfect character of the available CSIT and prescribed QoS constraints.

### A. *Scenario 1: homogeneous traffic*

Without loss of generality, Scenario 1 considers a set of $N_m = 8$ users all receiving information streams of the same traffic class, assumed to be RT traffic streams with a maximum allowable delay of $\breve{D}_m = 50$ ms. For the simulations shown here, we concentrate on the use of the MLWDF scheduler.

Figures 2 and 3 depict the attained average throughput as a function of the incoming traffic for CRA and DRA schemes, respectively, and for a variety of CSIT conditions under UPA and APA. Two common traits to all curves in these figures are: 1) for low incoming traffic, the system is able to dispatch it all, thus the throughput *vs* arrival curves initially increase with a slope of 1 up to the point where saturation appears, an indication that the system can no longer serve all the incoming traffic, and 2) it can be clearly appreciated how regardless of the design and power allocation, better CSIT quality (i.e., higher $\zeta$) leads to improved system performance.

Concentrating now on the CRA case (Fig. 2), it is very remarkable the great advantage offered by the robust design in front of the naive design. The poor performance of the naive design even for accurate CSIT ($\zeta = 0.99$) is to be sought in the continuous nature of the system. Figure 4 presents *pdf* estimates of the equivalent channel gains used to perform the resource allocation when using the robust design ($F^{-1}_{\delta_{b,m}|\boldsymbol{\Phi}_{b,m}}(\breve{\pi}_m)$) for poor ($\zeta = 0.9$) and accurate ($\zeta = 0.99$) CSIT, and also when using the naive design ($\bar{\delta}_{b,m}(t)$). Notice that only one pdf is show for the naive

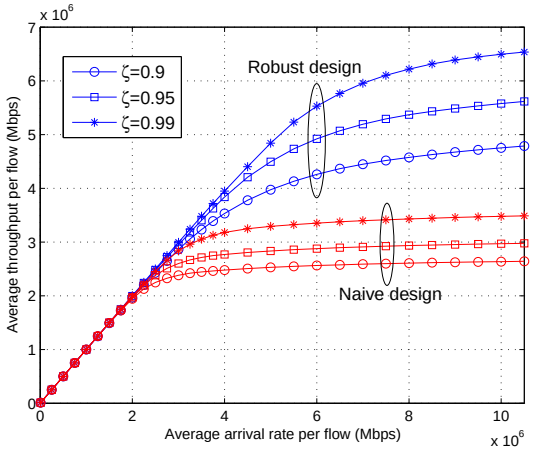


Figure 2: Scenario 1. Continuous rate allocation.

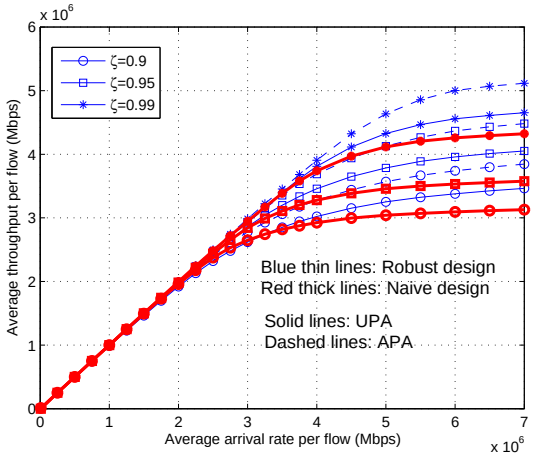


Figure 3: Scenario 1. Discrete rate allocation.

design as this does not take into account the imperfect character of the CSIT. Note how the channel gains onto which the robust design performs the resource allocation do indeed vary in accordance with the CSIT quality. In particular, when CSIT is poor the robust design utilizes conservative estimates of the channel gains (mostly between 0 and 2) owing to the uncertainty with which they have been estimated. In contrast, when CSIT is accurate the effective channel gains are larger to reflect the larger confidence in the estimation. Note that as $\zeta \to 1$, both designs, naive and robust, will lead to equal effective channel gains, which in turn result in the same *pdf*. It may seem somewhat counterintuitive that for accurate CSIT, the robust detector so clearly outperforms the naive design. The explanation is due to the fact that even for an extremely

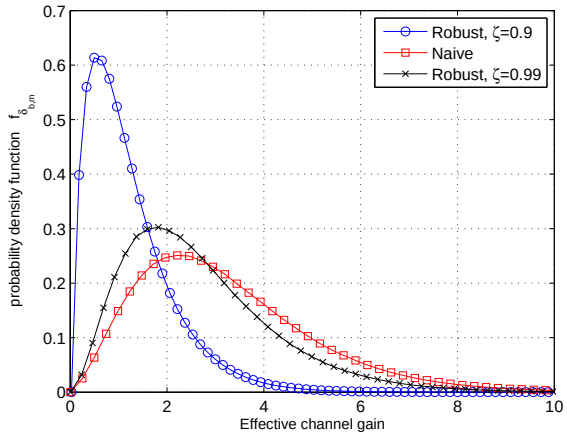


Figure 4: Scenario 1. Probability density function (*pdf*) estimates of the equivalent channel gains.

accurate CSIT, when the presumed channel gain guiding the resource allocation process is almost identical to the actual channel gain, there is always a $\frac{1}{2}$ probability that the presumed gain exceeds the real gain, in which case the capacity for that particular frequency band drops to zero (capacity outage). In fact, it can be observed in Fig. 2 that for $\zeta = 0.99$, the achievable saturated throughput for the naive design actually approaches half the one of the robust design due to this effect. The performance of the naive design could certainly be improved if a certain *safety rate backoff* [18] was applied when estimating the channel gain employed to drive the resource allocation. In fact, this is the mechanism that actually implements the robust design introduced in this paper, with the advantage that it is optimally tuned to the available CSIT accuracy and the specific QoS requirements (i.e., outage BLER). Finally, it is worth commenting that, in the CRA case, results are only shown for UPA as the benefits of APA were found to be virtually negligible (see [11]).

Turning now the attention to the DRA case (Fig. 3), it can be observed that the robust design still significantly outperforms the naive design. Nonetheless, the performance gap between them is far less pronounced than in the CRA case because indeed, operating with a discrete modulation and coding set (MCS) where a given MCS is optimum over an SNR range, somehow induces the *safety rate backoff* that was mentioned before. It is also remarkable that APA (for clarity shown only for the robust design) results in significant throughput gains, thus indicating that this strategy plays an important role in systems with constrained MCS sets.

Complementing the throughput plots, Fig. 5 presents the average BLER results for the robust and naive designs under CRA and DRA for two different levels of CSIT accuracy. Remarkably,

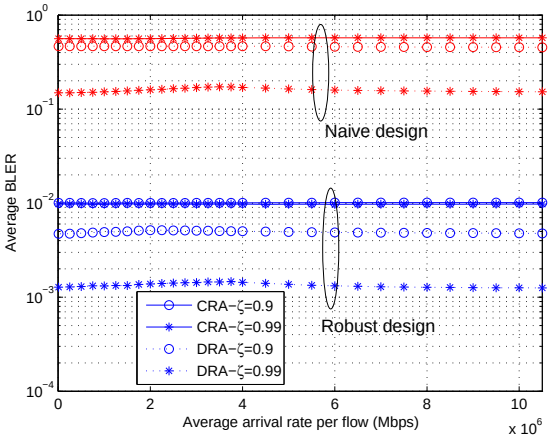


Figure 5: Scenario 1. Average BLER for robust and naive designs for $\zeta = 0.9$ and $\zeta = 0.99$ under CRA and DRA.

when using CRA the robust design satisfies with equality the prescribed instantaneous BLER outage constraint independently of the CSIT quality, thus somehow highlighting the optimum nature of the CRA scheme. When using the robust design in combination with DRA, the resulting average BLER is below the target as a result of often having to select the transmission mode in a rather conservative fashion due to the lack of granularity in the MCS set. In sharp contrast with the robust scheme, the naive design, regardless of the rate allocation and CSIT accuracy, is found unable to fulfill the BLER outage constraint.

### *B. Scenario 2: heterogeneous traffic*

Scenario 2 is defined by a collection of users with different traffic requirements. In particular, it is assumed that there are $N_m^{\text{RT}}$ real-time traffic users, $N_m^{\text{nRT}}$ non real-time users and $N_m^{\text{BE}}$ best effort users with maximum allowable delays of 50, 100 and 300 ms, and with delay outage probabilities of 0.01, 0.1 and 0.1, respectively. Without loss of generality, for the results shown here it is assumed that $N_m^{\text{RT}} = N_m^{\text{nRT}} = N_m^{\text{BE}} = 4$ users, thus totalling 12 different users in the system. Furthermore, this section assumes the utilization of the robust design in combination with DRA+APA, as this configuration is the one most likely to be used in practical systems. Figure 6 evaluates how CSIT inaccuracies impact the different traffic classes for the case of MLWDF scheduling. First note that the throughput for the different traffic classes behaves in the expected way. In particular, it is clearly observed how, as the traffic arrivals increase, the

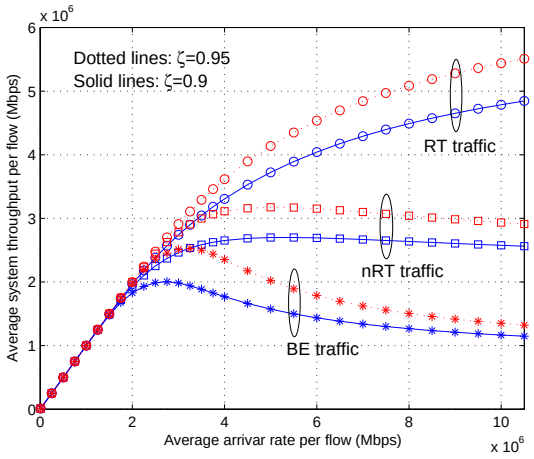


Figure 6: Scenario 2. Throughput under heterogeneous traffic conditions. MLWDF scheduling. Robust design (DRA+APA).

performance of the BE and nRT classes is sacrificed in order to keep serving the RT traffic class, which has the most stringent QoS requirements. Also, as expected, the more degraded the CSIT quality is, the lower the attained throughputs for all user types.

Figure 7 compares the throughput performance of different scheduling policies (PF, MLWDF and EXP) when considering a CSIT quality defined by $\zeta = 0.95$. Focusing first on the results for MLWDF and EXP schedulers, it can be concluded that both strategies are appropriate to handle heterogeneous traffic requirements despite notable differences between them: the EXP scheduler tends to favour the RT and nRT users at the cost of hardly providing any services to BE users. In fact, note how BE users, after a certain traffic arrival rate, stop being serviced under EXP scheduling and when the service to the RT class shows the first signs of saturation (around 8 Mbps), service to nRT users is drastically reduced. In contrast, the MLWDF, at least up to a certain traffic arrival rate, aims at splitting resources among all traffic classes. For comparison purposes, results for the PF scheduler are also shown. Notice how this scheduler allocates resources among users without taking into account the traffic class being served, thus resulting in similar throughput for all users in the system. Obviously, such scheduling policy cannot be considered adequate to handle heterogeneous traffic requirements. Figure 8 depicts the service coverage[8] for the different schedulers where, as in previous figures, a CSIT quality

[8]Defined as the percentage of users who achieve their QoS requirements in terms of minimum throughput or maximum allowable average or absolute delay.

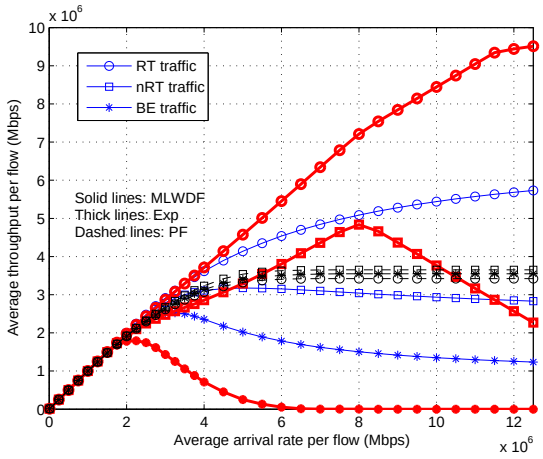


Figure 7: Scenario 2. Throughput performance under heterogeneous traffic conditions for different schedulers. Robust design with DRA+APA.

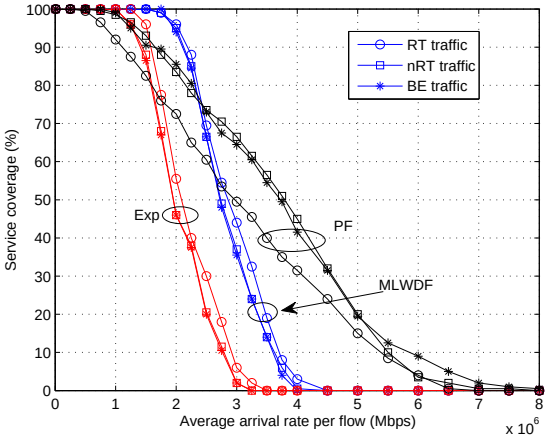


Figure 8: Scenario 2. Service coverage under heterogeneous traffic conditions for different schedulers. Robust design with DRA+APA.

defined by $\zeta = 0.95$ has been considered. Service coverage reveals that all schedulers are able to guarantee the QoS restrictions up to an scheduler-dependent limit, where from this point onwards the service coverage drops dramatically. As expected, the PF scheduler for both RT and nRT traffic, is the most sensitive one to an increase in traffic arrivals.

It is worth mentioning at this point that the results shown here are just a (representative) small fraction of the ones that can be shown. Measures such as the Jain's fairness index for delay and throughput, probability of exceeding the prescribed delay and many others can be derived from

the proposed robust design. It should be clear from the results so far that a single metric is usually not representative of the merits or weaknesses a particular choice of system parameters may have, in fact, several performance indicators are often required to evaluate it. Remarkably, the proposed framework can be used to effectively evaluate the overall system performance by realizing that the only computationally-intensive part of the framework (i.e., the statistical characterization of the SINR under imperfect CSIT) needs only be evaluated once. This physical layer characterization can then serve to estimate the performance of different rate allocation strategies (DRA or CRA), power allocation techniques (APA or UPA), scheduling policies or arrival rates, to name a few, and see how the CSIT accuracy affects the different performance metrics.

### *C. Scenario 3: correlation effects*

The last considered scenario is devoted to evaluate the effects of antenna correlation on the system performance using as a baseline the same setup of Scenario 2. For the results shown here, MLWDF scheduling in combination with DRA+APA is assumed.

Figures 9 and 10 present results for throughput and delay, respectively, when a significant amount of correlation is present at the transmitter or receiver side ($\rho_{tx} = 0.5$ or $\rho_{rx} = 0.5$). Figure 9 reveals that transmit correlation ($\rho_{tx} = 0.5$) has hardly any effect on the throughput and it can even bring along (rather minor) performance improvements, an effect well described in the literature (see Chapter 5 in [49]) in the context of MRT schemes with partial CSIT. In contrast, receiver correlation does indeed induce a significant performance loss that can be seen to affect every single traffic class, specially when the system reaches saturation. Similar conclusions can be drawn regarding the delay based on the results shown in Fig. 10, where it can be observed that receive correlation leads to larger delays whereas transmit correlation can even be marginally beneficial.

## VI. CONCLUSIONS

This paper has considered the problem of resource allocation and scheduling in the context of (downlink) spatially correlated MIMO-OFDMA networks where the available CSIT information has limited accuracy. An analytical cross-layer framework has been developed that allows design decisions such as power, subcarrier and rate allocation, and scheduling strategy, to be taken in a robust manner by incorporating the degree of CSIT accuracy while maximising some

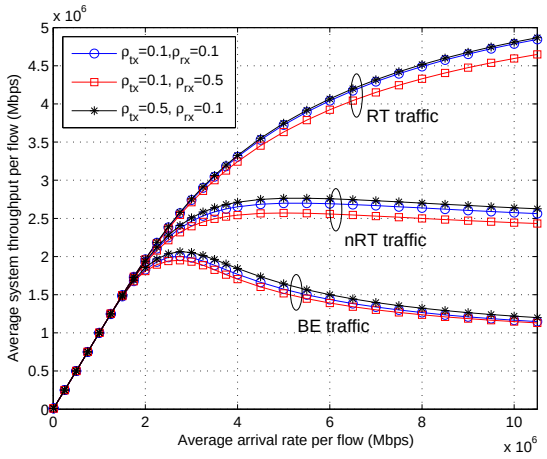


Figure 9: Scenario 3. Throughput performance under heterogeneous traffic conditions for different transmit correlation conditions. Robust design with DRA+APA.

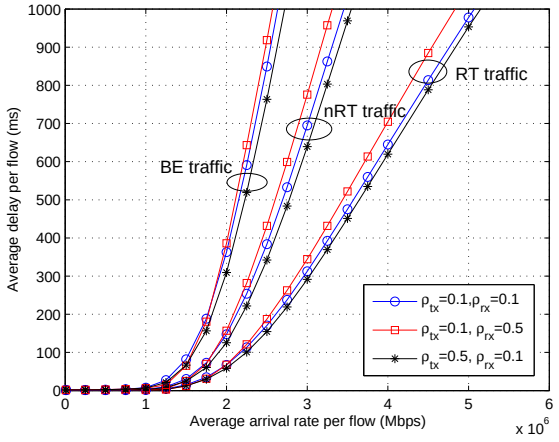


Figure 10: Scenario 3. Delay performance under heterogeneous traffic conditions for different transmit correlation conditions. Robust design with DRA+APA.

utility function subject to prescribed QoS constraints. The presented framework allows the study of many performance metrics of practical relevance such as throughout, fairness (in the form of the JFI), delay or service coverage. Numerical results drawn using parameters from current 4G systems serve to illustrate the versatility of the proposed cross-layer design in studying different SRA techniques. Transmit spatial correlation has been shown to play a minor role in overall network performance whereas receive correlation has indeed been observed to significantly impact the different performance metrics. Numerical results have clearly shown the

performance advantages of the proposed robust framework when compared to a naive design that neglects the imperfect nature of the CSIT, thus reveling the importance of incorporating statistical CSIT characterizations in the cross-layer SRA design. Further research work is expected to progress along three different, but rather intertwined, lines: 1) the generalization of the current framework to allow per-user multi-stream transmission, 2) the incorporation of multi-user MIMO mechanisms, and 3) the extension of the framework to multicell scenarios.

## ACKNOWLEDGEMENTS

This work has been supported in part by the MEC and FEDER under project AM3DIO (TEC2011-25446), Ministerio de Economia y Competitividad (MINECO), Spain.

## REFERENCES

[1] I. Fu, Y. Chen, P. Cheng, Y. Yuk, Y. Kim, and J. Kwak, "Multicarrier technology for 4G WiMax system," *IEEE Communications Magazine*, vol. 48, no. 8, pp. 50–58, 2010.

[2] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-advanced: next-generation wireless broadband technology," *IEEE Wireless Communications*, vol. 17, no. 3, pp. 10–22, 2010.

[3] M. Alasti, B. Neekzad, J. Hui, and R. Vannithamby, "Quality of service in WiMAX and LTE networks," *IEEE Communications Magazine*, vol. 48, no. 5, pp. 104–111, 2010.

[4] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part I: theoretical framework," *IEEE Tran. Wireless Commun.*, vol. 4, no. 2, pp. 614–624, 2005.

[5] ——, "Cross-layer optimization for OFDM wireless networks-part II: theoretical framework," *IEEE Tran. Wireless Commun.*, vol. 4, no. 2, pp. 625–634, 2005.

[6] D. Hui, V. Lau, and W. Lam, "Cross-layer design for OFDMA wireless systems with heterogeneous delay requirements," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 2872–2880, 2007.

[7] C. Mohanram and S. Bhashyam, "Joint subcarrier and power allocation in channel-aware queue-aware scheduling for multiuser OFDM," *IEEE Transactions on Wireless Communications*, vol. 6, no. 9, pp. 3208–3213, 2007.

[8] Z. Kong, Y.-K. Kwok, and J. Wang, "A Low-Complexity QoS-Aware Proportional Fair Multicarrier Scheduling Algorithm for OFDM Systems," *IEEE Trans. Vehic. Technol*, vol. 58, no. 5, pp. 2225–2235, 2009.

[9] G. Song, Y. Li, and L. Cimini, "Joint channel- and queue-aware scheduling for multiuser diversity in wireless OFDMA networks," *IEEE Tran. Commun.*, vol. 57, no. 7, pp. 2109–2121, 2009.

[10] N. Zhou, X. Zhu, Y. Huang, and H. Lin, "Low complexity cross-layer design with packet dependent scheduling for heterogeneous traffic in multiuser OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 6, pp. 1912–1923, 2010.

[11] G. Femenias, B. Dañobeitia, and F. Riera-Palou, "Unified approach to cross-layer scheduling and resource allocation in OFDMA wireless networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, pp. 1–19, 2012.

[12] R. Wang and V. Lau, "Cross layer design of downlink multi-antenna OFDMA systems with imperfect CSIT for slow fading channels," *IEEE Trans. on Wireless Communications*, vol. 6, no. 7, pp. 2417–2421, 2007.

[13] I. Wong and B. Evans, "Optimal resource allocation in the OFDMA downlink with imperfect channel knowledge," *IEEE Transactions on Communications*, vol. 57, no. 1, pp. 232–241, 2009.

[14] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 288–296, 2009.

[15] M. Awad, V. Mahinthan, M. Mehrjoo, X. Shen, and J. Mark, "A dual-decomposition-based resource allocation for OFDMA networks with imperfect CSI," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 5, pp. 2394–2403, 2010.

[16] R. Aggarwal, M. Assaad, C. Koksal, and P. Schniter, "Joint scheduling and resource allocation in the OFDMA downlink: Utility maximization under imperfect channel-state information," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5589–5604, 2011.

[17] C. C. Zarakovitis, Q. Ni, D. E. Skordoulis, and M. G. Hadjinicolaou, "Power-efficient cross-layer design for OFDMA systems with heterogeneous QoS, imperfect CSI, and outage considerations," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 2, pp. 781–798, 2012.

[18] Y. Hu and A. Ribeiro, "Optimal wireless multiuser channels with imperfect channel state information," in *IEEE ICASSP*, 2012.

[19] G. Femenias and F. Riera-Palou, "Cross-layer resource allocation and scheduling in LTE systems under imperfect CSIT," *IEEE/IFIP Wireless Days*, November 2013.

[20] *Third Generation Partnership Project, Technical Specification Group Radio Access Network; Physical layer aspects for evolved Universal Terrestrial Radio Access (UTRA)*. 3GPP Std. TR 25.814 v. 7.0.0, 2006.

[21] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*. 3GPP Std. TS 36.211. Release 8, 2008.

[22] *Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and Channel Coding*. 3GPP Std. TS 36.212. Release 8, 2008.

[23] C. Cox, *An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications*. Wiley, 2012.

[24] X. Wang, G. Giannakis, and A. Marques, "A unified approach to QoS-guaranteed scheduling for channel-adaptive wireless networks," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2410–2431, 2007.

[25] J. Kermoal, L. Schumacher, K. Pedersen, P. Mogensen, and F. Frederiksen, "A stochastic MIMO radio channel model with experimental validation," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 6, pp. 1211–1226, 2002.

[26] J. G. Proakis, *Digital Communications*, 4th ed. McGraw Hill, 2001.

[27] X. Zhang, D. P. Palomar, and B. Ottersten, "Statistically robust design of linear MIMO transceivers," *IEEE Tran. Signal Processing*, vol. 56, no. 8, pp. 3678–3689, 2008.

[28] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei, "An overview of MIMO communications-a key to gigabit wireless," *Proc. IEEE*, vol. 92, no. 2, pp. 198–218, 2004.

[29] T. Lo, "Maximum ratio transmission," *IEEE Trans. Commun.*, vol. 47, no. 10, pp. 1458–1461, 1999.

[30] M. R. McKay, P. J. Smith, and I. B. Collings, "New properties of complex noncentral quadratic forms and bounds on MIMO mutual information," in *IEEE International Symposium on Information Theory (ISIT)*, 2006, pp. 1209–1213.

[31] W. Conradie and C. Troskie, "The exact non-central distribution of a multivariate complex quadratic form of complex normal variables," *South African Statistical Journal*, vol. 18, pp. 123–134, 1984.

[32] M. R. McKay and I. B. Collings, "General capacity bounds for spatially correlated Rician MIMO channels," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3121–3145, 2005.

[33] T. Ratnarajah and R. Vaillancourt, "Quadratic forms on complex random matrices and multiple-antenna systems," *IEEE Transactions on Information Theory*, vol. 51, no. 8, pp. 2976–2984, 2005.

[34] A. A. Mohsenipour, "On the distribution of quadratic expressions in various types of random vectors," PhD Thesis, University of Western Ontario, 2012.

[35] A. M. Mathai and S. B. Provost, *Quadratic forms in random variables: theory and applications*. Marcel Dekker, Inc., 1992.

[36] A. Annamalai, C. Tellambura, and J. Matyjas, "A new twist on the generalized Marcum Q-function $Q_M(a, b)$ with fractional-order $M$ and its applications," in *IEEE Consumer Communications and Networking Conference (CCNC)*, 2009, pp. 1–5.

[37] J. Marcum, "Table of $Q$ functions, U.S. Air Force Project RAND Res. Memo. M-339, ASTIA Document AD 1165451," Rand Corporation, Santa Monica, CA, Tech. Rep., 1950.

[38] I. C. Wong and B. Evans, *Resource allocation in multiuser multicarrier wireless systems*. Springer, 2008.

[39] A. J. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge University Press, 2005.

[40] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *The Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[41] S. Shakkottai and A. Stolyar, *Scheduling algorithms for a mixture of real-time and non-real-time data in HDR*. Bell Laboratories, Lucent Technologies, 2000.

[42] M. Andrews, S. Borst, F. Dominique, P. Jelenkovic, K. Kumaran, K. Ramakrishnan, and P. Whiting, "Dynamic bandwidth allocation algorithms for high-speed data wireless networks," *Bell Labs Technical Journal*, vol. 3, no. 3, pp. 30–49, 1998.

[43] S. Shakkottai and A. Stolyar, "Scheduling for multiple flows sharing a time varying channel: the exponential rule," *Bell Labs Technical Report*, 2000.

[44] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Tran. Commun.*, vol. 54, no. 7, pp. 1310–1322, 2006.

[45] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[46] "LTE - evolved universal terrestrial radio access (E-UTRA) physical channels and modulation - (3GPP TS 36.211 version 8.7.0 release 8)," 3GPP, Tech. Rep., 2009.

[47] "LTE - evolved universal terrestrial radio access (E-UTRA) base station (BS) radio transmission and reception - (3GPP TS 36.104 version 8.8.0 release 8)," 3GPP, Tech. Rep., 2009.

[48] "LTE - evolved universal terrestrial radio access (E-UTRA) physical layer procedures - (3GPP TS 136.213 version 8.8.0 release 8)," 3GPP, Tech. Rep., 2009.

[49] H. Bolcskei, D. Gesbert, C. B. Papadias, and A.-J. v. d. Veen, *Space-Time Wireless Systems: From Array Processing to MIMO Communications*, 1st ed. New York, NY, USA: Cambridge University Press, 2008.