Dynamic Resource Allocation for Time-varying Channels in Next Generation Cellular Networks Part I: A Mathematical Framework

Tania Villa*, Raymond Knopp*, Ruben Merz‡

∗ Mobile Communications Dept., Eurecom, Sophia Antipolis, France

‡ Swisscom, Group Strategy & Innovation, Switzerland

Abstract

Next generation cellular networks will present challenging interference scenarios due to the difficulty of centralized planning and continued support of mobility. In this context, modeling and analytical study is still required for time-varying inter-cell interference and imperfect resolution of the channel state at the transmission end. In this paper, we first present a mathematical framework using information-theoretic quantities that can be applied to the analysis of heterogeneous networks and provide insight into the design of resource scheduling policies. Specifically, we consider the problem of variable resource allocation for IR-HARQ schemes across time-varying channels, arising from either fading with unknown or partial channel state information, time-varying interference, or a combination of both. With our framework, we avoid the need for extensive simulations and can flexibly address the development of resource allocation policies with and without constraints on the outage probability, which to a first degree represents the latency of the protocol. The policies are distributed, applicable for uplink and downlink, and based on the dynamic adaptation of the physical dimensions across HARQ rounds. Our results show a significant gain from adapting the resources across rounds, and we identify specific cases where it provides the highest gain when compared to fixed-allocation schemes.

Index Terms

Heterogeneous networks, resource allocation, rate adaptation, LTE, HARQ, incremental redundancy

I. Introduction

Traditionally, hybrid automatic repeat request (HARQ) has been used to recover from transmission errors, therefore decreasing the probability of unsuccessful decoding. In incremental redundancy (IR)

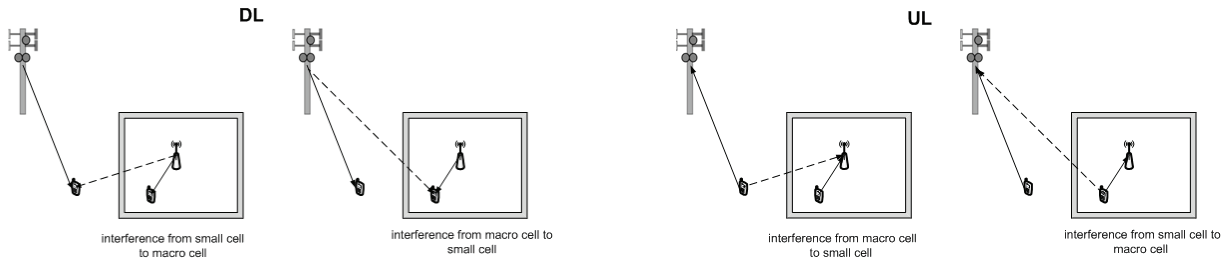
HARQ, the retransmission consists of the addition of new parity bits [1]. IR refers to the different puncturing patterns applied by the physical layer to the original codeword transmission and retransmissions.

Another common technique to improve system performance is to vary the coding rate across the retransmissions to adapt to and exploit channel and traffic variations. The code rate can be fine-tuned by puncturing. In the 3GPP Long Term Evolution (LTE), for instance, the code rate and rate matching, together with the number of resources allocated for one transmission determine the transport-block size [2]. In essence, rate adaptation tailors the modulation and coding scheme (MCS) to the current channel conditions, which determines the link data rate or error probability. Because the MCS represents the combination of a modulation scheme and a coding rate, in terms of spectral efficiency, it specifies the number of information bits per modulation symbol. The use of rate adaptation provides manufacturers an incentive to implement more advanced receivers since those receivers will result in higher end-user data rates than standard receivers [3].

In LTE, partial or outdated channel state information (CSI) can occur because of moderate to high mobility, of insufficient uplink channel quality information (CQI) periodicity or of non-stationary inter-cell interference. The latter will become more and more important with LTE release 10 networks and their inherent heterogeneity. Hence, the scheduler must operate blindly for rate adaptation and can only benefit from feedback after the first HARQ transmission round in the form of ACK/NACK signaling.

Heterogeneous networks (HetNets) implies there are different types of cells in the network, i.e. low power base stations are distributed throughout a macrocell network. These low power base stations can be microcells, picocells, relays, femtocells or distributed antenna systems [4]. On the one hand, microcells, picocells and relays are deployed by the operator to increase the capacity and coverage in public places, enterprises buildings, etc. On the other hand, femtocells are user-deployed at home to improve capacity. We generally denote low power base stations by small cells.

Cellular HetNets typically operate on licensed spectrum owned by the network operator. The most severe interference is experienced when the small cells are deployed on the same frequency carrier as the macrocells [5]. LTE, in its release 10, identifies more challenging interference scenarios since interference can come across layers (macro–small cell, small–macrocell), for example, a macrocell user far from the base station is transmitting at a very high power hurting the small cells in the vicinity. Interference can also be experienced between small cells in both the uplink (UL) and downlink (DL) channels (see figure 1). In the case of inter-layer interference, the macrocell scheduler has to take into account the bursty interference



(a) Interference scenarios DL. (b) Interference scenarios UL.

Fig. 1. Figure (a) shows the interference scenarios for HetNets in the DL, figure (b) shows the interference scenarios for HetNets in the UL

from the small cells since they will be serving only a couple of users. Given the fact that in HetNets there is no controller managing the allocation process [6], operators will not be able to handle interference between small cells in a centralized manner (centralized frequency planning). The design of distributed algorithms and techniques allowing for an efficient utilization of infrastructure is one of the key challenges in HetNets [7]. In this type of approach, the small cell adapts its performance independently from other cells avoiding the need for any *a priori* centralized frequency planning and without having to exchange information or sending it to a central controller. They rely only on feedback, avoiding uncontrolled delays.

The remainder of this paper is organized as follows. We begin by describing the related work and our contributions in section II. The signal and system model are presented in section III. Analytical expressions for throughput are derived in section IV. Section V presents a motivating example for rate adaptation with and without interference. Our resource allocation policies for practical systems are exposed in section VI. Finally, we conclude in section VII.

II. Related work and contributions

Extensive research has explored adaptive techniques. However, very little attention has been paid to the more performance-limited case of interference. Early work in [8] suggests a gain from adaptive policies. By deriving the Shannon capacity regions of variable rate and power, it is shown that the maximum capacity is achieved when the rate is varied based on the channel variations. More recently, [9] explores rate adaptation with successive interference cancellation receivers for Multiple-Input Multiple-Output (MIMO) systems with outdated channel state information and Gaussian signals.

The throughput of HARQ has been investigated for Gaussian input signals [10] over a Gaussian channel with fading and in the limit of infinite block length. In [11], the long-term throughput analysis of a HARQ protocol under slow-fading channels is presented for fixed-rate, variable-power transmissions under the framework of the renewal-reward theory of [10]. Rate adaptation for HARQ protocols under

delay constraints is studied in [12], and for time-correlated channels in [13] and [14]. In [15], rate and transmit power are adapted under perfect CSI. Power adaptation is also presented in [16] to minimize the outage probability and in [17], both power and rate control are derived through dynamic programming without outage constraints. Combined power and rate adaptation is also presented in [18], and in [19], the optimization of either the packet drop probability or the average transmit power is shown for the case of IR-HARQ with a maximum number of retransmissions. In [20], the information-theoretic approach of [10] is adapted to variable rate transmissions in the case of HARQ with IR. The idea of changing the MCS for retransmissions is presented in [21], for IP video surveillance camera traffic by assigning additional redundancy to the retransmissions and reducing the estimated CQI.

Recent so-called *rateless or fountain* coding techniques with IR for additive-noise channels are also reported in [22], [23], [24], [25]. When combined with a HARQ link-layer protocol, these coding schemes allow for transmission over unknown channels without the need for sophisticated rate adaptation policies, and whose instantaneous rate (or spectral-efficiency) depends on the time the decoder is able to decode the message. The basic principle for this type of transmission was introduced for content distribution over the internet and broadcast networks by Luby [26] using so-called *LT-codes* for erasure channels. These were improved by Shokrollahi with his invention of *Raptor codes* [27]. The latter were then adapted for AWGN channels in [28].

All of these coding strategies are structured, and, in particular Perry *et al*'s *Spinal Codes*, can approach Shannon's AWGN channel capacity with varying degrees of encoding and decoding complexity provided the number of transmissions is allowed to grow without bound. Although not shown in [22] [24] it may very well be true for any ergodic time-varying additive-noise channel. An extension of the promising superposition coding technique designed for successive decoding at the receiver considered by Erez *et al* was also described for time-varying channels without an *a priori* stochastic model [22]. This considered the performance of their rateless coding construction for a small number of transmission rounds.

In this work, we consider similar rateless strategies for time-varying channels for a finite and small number of transmission rounds, potentially allowing for a residual outage probability after the maximum number of rounds. Imposing a quasi-finite duration for transmission is often required to minimize latency in data transmission networks. For instance, the HARQ protocol of LTE reference channels [29] is tuned to offer an approximate 1% outage rate after two transmission rounds which allows for a one-way latency of 10ms for 99% of transmissions. This can, of course, be tuned to offer different latency-throughput

tradeoffs. Since the maximum number of transmission rounds is fixed in such protocols, it seems natural that the number of dimensions used in each round should be optimized in order to maximize throughput, by progressively decreasing code rate across rounds. We should note that the latter is not a requirement in rateless coding with an unbounded number of transmission rounds.

A. Contributions

In this work, we develop and evaluate dynamic resource allocation policies for IR-HARQ schemes under the presence of interference. We consider resource allocation both via rate (e.g. MCS) and physical dimensions adaptation. Rather than performing extensive simulations which are left to the subsequent part of this work, we focus here on deriving analytical expressions for the long-term throughput and consider cases with and without a constraint on the outage probability, which is a first order representation of the latency of the protocol. We address practical cases with outage at the end of the retransmission protocol (which is taken care of by upper layer retransmission protocols). Our policies are distributed and do not depend on a centralized resource allocation. They can be used for a macro base station or a small cell scheduler indifferently. Our contributions are the following:

- We motivate the use of inter-round resource allocation through a simple but illustrative analysis with Gaussian signals with and without interference.
- We provide a mathematical framework for the analysis of HetNets. Under this framework, we derive analytical expressions based on mutual information modeling, that capture the throughput performance.
- We develop distributed dynamic resource allocation policies that are applicable both for the UL and DL channels. Our policies are based on the dynamic adaptation of the physical dimensions and coding rate used in each HARQ round. The latter is a real possibility in schedulers for LTE base stations and, to the best of our knowledge, no well-known methodology exists for optimizing the resource allocation across transmission rounds for time-varying channels. We analyze such an optimization when time-variation arises from either fading with unknown or partial channel state information, time-varying interference, or a combination of both. Moreover, the resource allocation policies that can be developed based on the conclusions of this work could be applied to other coding strategies such as those proposed in [22] and [23]. The latter would require appropiate link-layer HARQ protocols (e.g. [30]) adapted to such dynamic rateless coders. In our subsequent work [31], we study the

benefits of such dynamic resource scheduling in the context of the LTE HARQ protocol combined with standardized coded-modulation. This combination of the fixed-rate turbo-code with dynamic resource allocation and the rate-matching permutation amounts to doing rateless coding over a small number of transmission rounds (2-4).

III. Signal and System Model

We consider a slotted transmission scheme and we take an information-theoretic approach to analyze the throughput performance. When there is more than one user, we assume that all transmissions in every slot are synchronized and we randomize the interference process with the use of activity factors. The latter models sporadic interference patterns characteristic of future heterogeneous networking deployments, in particular the interference seen from small cell base stations with bursty traffic in the receiver of a macrocell user. It can also model dual-carrier networks with cross-carrier scheduling. In this type of network, we can talk about clean and dirty carriers. On the one hand, clean carriers are used by the macrocell to carry their data plus signaling for small cells because of their controlled interference property. On the other hand, dirty carriers are interfering carriers where the "cleaning" is done with the use of HARQ.

We consider a maximum of $\mathbb{M}_{max}$ HARQ transmission rounds and the channel is either independent and identically distributed (iid) or constant over all the transmission rounds of the protocol. After each transmission we receive an error-free acknowledgment (ACK or NACK) indicating a successful or unsuccessful transmission. We define the probability of outage as being unsuccessful to correctly receive the information at the end of the HARQ protocol. This probability translates to the latency of the protocol and quality of service (QoS) in our system.

In general, we define R_r as the code rate at the $r\mathrm{th}$ round. For a particular user, we define the number of dimensions in time as T_{dim} and the number of dimensions in frequency as L_r. Let L'_r be the number of dimensions in frequency up to round r. Then, at each transmission round, the total number of dimensions is $L'_r T_{dim}$. Assuming the channel does not vary during T_{dim} time dimensions and for a packet length of B information bits, the rate R_r at the rth round, in bits/dim is given by:

$$R_r = \frac{\log_2 B}{L'_r T_{dim}} \quad \text{bits/dim.} \tag{1}$$

In IR-HARQ, the retransmission consists of the same set of information bits as the original, however, the set of coded bits are chosen differently and they may contain additional parity bits. In each of the

transmission rounds there are $L_r T_{dim}$ dimensions, however, this number is not necessarily the same across rounds according to the LTE standard [32] (see figure 2).

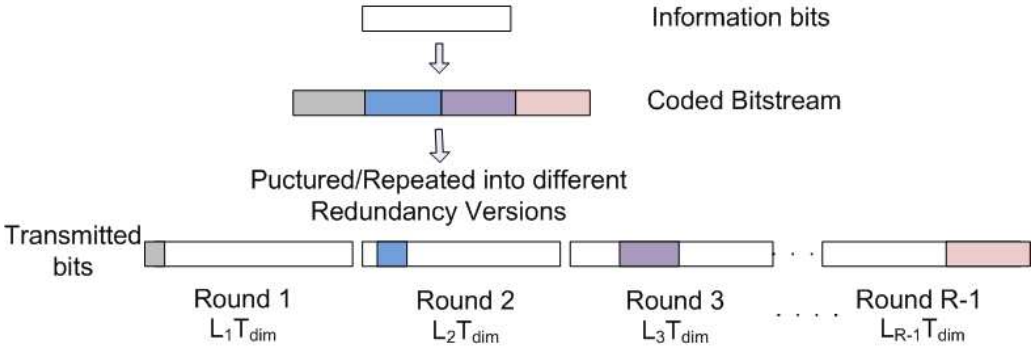


Fig. 2. Coding Model

In the context of LTE, the number of physical dimensions $L_r T_{dim}$ refers to the number of resource blocks allocated to one user in one subframe of 1 ms duration, i.e. one Transmission Time Interval (TTI). There are at most two transport blocks delivered to the physical layer in the case of spatial multiplexing [33]. In a single-user LTE system, there is only one transport block in one TTI, representing only one codeword "in the air" at the same time. Each transport block is carried by an HARQ process, and each process is assigned to a subframe (number of processes is fixed). In our model, if the number of dimensions for a user is less than the maximum number of available resources N_T, $(L_r T_{dim} < N_T)$, then the rest will not be utilized. Although not possible in the current LTE standard, one could propose to assign the unused resources to transmit multiple codewords in parallel (at the same time), to increase the throughput. In a multiuser system, the remaining dimensions would be allocated to other users and thus the efficiency of the protocol should be chosen to maximize the aggregate spectral efficiency of the cell.

We consider N_u transmitters, where user 0 is the transmitter of interest, and the remaining $N_u - 1$ transmitters are interferers. We model an OFDMA physical layer with K subcarriers. We let $x_{j,k}$ be the input signal of the jth user on the kth subcarrier, $\mu_{j,k}$ the activity factor, and P_j the transmission power. We assume discrete signals with equal probabilities and size of the constellation $\mathcal{S}$, z_k is the zero mean complex Gaussian noise with variance σ^2. Since we assume Rayleigh fading, $h_{j,k}$ is a circularly symmetric complex Gaussian random variable with unit mean. The received signal y is given by:

$$y = \sum_{j=0}^{Nu-1} \sqrt{P_j} \sum_{k=0}^{K-1} h_{j,k} \mu_{j,k} x_{j,k} + z_k \tag{2}$$

Variations in the channel are caused at the receiver because of the activity factor plus the frequency

shifting from the resource allocation process. For the interfering users, the channel variation depends on whether we consider the UL or DL. In the UL, it is caused by the interference coming from different user terminals. In the DL, the activity factors will introduce variations originated from the fact that the interfering cells are not active the whole time.

Under this mathematical framework we consider cases with and without an outage constraint which translates to the latency of the protocol. For the cases without an outage constraint, we can talk about a residual outage probability which we assume is handled by an upper layer ARQ process on top of HARQ [34]. We can relate the rate adaptation policy to a rate quantization process where the rates are quantized into equal-size bins. The bins are defined by their probability of occurrence, and the chosen rate in each bin should be the rate for which any channel falling in the bin should allow correct decoding and will maximize the overall throughput. Moreover, the quantization steps are refined as transmission rounds progress.

IV. Information-theoretic quantities

We target LTE release 10 networks with an OFDMA physical layer, and we study both the single-user and one dominant interferer cases. For the sake of analytical tractability, we show the derivations focusing on one subcarrier, with unitary power and distance, and we drop the indexes.

Let x be the input signal from a discrete distribution with equal probabilities and size of the constellation $\mathcal{S}$, z is the zero mean complex Gaussian noise with variance σ^2, h is a circularly symmetric complex Gaussian distributed random variable with unit mean. Then, the received signal y is given by: $y = hx + z$. The general expression for mutual information, in bits/dim, when the input signals come from discrete constellations (to model practical systems) with equally probable symbols is given by [35]:

$$I(Y;X|H=h) = \frac{1}{\mathcal{S}} \sum_{i=0}^{\mathcal{S}-1} \int_y f(y|x_i,h) \log_2 \left[\frac{f(y|x_i,h)}{\frac{1}{\mathcal{S}} \sum_k f(y|x_k,h)} \right] dy \tag{3}$$

In the case of one dominant interferer, the signal model is $y = h_1 x_1 + h_2 x_2 + z$ and the mutual information is given by:

$$I(Y;X|H=h) = \frac{1}{\mathcal{S}_1 \mathcal{S}_2} \sum_{x_1} \sum_{x_2} \int_y f(y|x_1,x_2,H) \log_2 \left[\frac{\sum_{x_2'} f(y|x_1,x_2',H)}{\frac{1}{\mathcal{S}_1} \sum_{x_1'} \sum_{x_2'} f(y|x_1',x_2',H)} \right] dy \tag{4}$$

Let H_r denote the vector of channel realizations in the rth round, then $I_r(H) = I_r(Y;X|H)$ denotes the corresponding instantaneous mutual information at round r. For IR-HARQ, mutual information is

accumulated over retransmissions. In the case of bursty interference, this permits some averaging of the fading and interference affecting the signal [36].

For a particular user, we define the mutual information at round r, in bits, as:

$$I_r(H) = T_{dim} \sum_{j=1}^{r} \sum_{k=1}^{L_j} I_{j,k}(H_j) \tag{5}$$

where $I_{j,k}(H_j)$ is the mutual information for the user at round j and subcarrier k, and it is given by (3), (4).

In the following sections, we refer to the mutual information in bits/dim. For this purpose, we define $I'_r(H_r)$ as the mutual information in bits/dim as:

$$I'_r(H_r) = \frac{1}{L'_r T_{dim}} I_r(H_r) \tag{6}$$

where L'_r is the number of dimensions up to round r, $\left(\sum_{j=1}^{r} L_j = L'_r\right)$.

Let P_{succ_1} be the probability of having a successful transmission in the first round, and $P_{succ_r, fail_{r-1}}$ the probability of not having a successful transmission in the $(r-1)$th round, but being successful in the rth round. Finally, let P_{out} represent the probability of outage at the end of the protocol. The overall throughput can thus be expressed as:

$$\bar{R} = P_{succ_1} R_1 + \sum_{r=2}^{\mathbb{M}_{max}} P_{succ_r, fail_{r-1}} \left(\frac{R_r}{r}\right) \quad \text{bits/dim.} \tag{7}$$

where the outage probability is given by $P_{out} = \Pr(r = \mathbb{M}_{max} + 1) = 1 - \sum_{r=1}^{\mathbb{M}_{max}} P_{succ_r}$.

In the case of an upper layer ARQ, the throughput expression becomes [10]:

$$\bar{R} = [1 - P_{out}] \left[P_{succ_1} R_1 + \sum_{r=2}^{\mathbb{M}_{max}} P_{succ_r, fail_{r-1}} \left(\frac{R_r}{r}\right) \right] \quad \text{bits/dim.} \tag{8}$$

Now, we can define the probabilities in (7) as a function of the mutual information:

$$P_{succ_r} = \Pr(I'_r(H_r) > R_r) \tag{9}$$

$$P_{succ_r, fail_{r-1}} = \Pr(I'_r(H_r) > R_r, I'_{r-1}(H_r) < R_{r-1}) \tag{10}$$

For a given channel realization h_r and a particular value of SNR, the maximum rate of reliable communication supported by the channel at round r is $I'_r(h_r)$ bits/s/Hz, which is a function of the random channel gain h_r and is therefore random. If the transmitter encodes data at a rate R_r bits/s/Hz, then at round r,

if the channel realization h_r is such that $I'_r(h_r) < R_r$, the transmission is called unsuccessful and this happens with probability $\Pr(I'_r(h_r) < R_r)$.

V. Modeling and Optimization of a Resource Scheduling Policy

In this section we provide some motivating examples of inter-round resource allocation with Gaussian signals with and without interference. The goal of these examples is to illustrate the benefits of dynamic resource allocation on simple analytical channels and the results should only be used as a first-order guideline for resource allocation in practical systems. In the next section, we explore the case for resource allocation in more practical scenarios. In a subsequent work [31], we apply some of these ideas for practical LTE scheduler design.

A. Initial analysis for interference-free networks

We start our analysis by looking into interference-free networks, i.e. we do not consider interference created by neighboring transmitters and we focus on single antenna systems, Single-Input Single-Output (SISO), although our model can be extended to MIMO. We denote by $\mathbb{M}_{max}$ the maximum number of transmission rounds. Let H_r denote the vector of channel realizations in the rth transmission round. Then $I\left(H_r\right)$ denotes the corresponding instantaneous mutual information. Accordingly, $I\left(H_1, \ldots, H_{\mathbb{M}_{max}}\right)$ defines the mutual information accumulated over $\mathbb{M}_{max}$ transmission rounds. In order to compute the mutual information, we assume Gaussian input signals (upper-bound on QAM modulation). For example, let us consider one subcarrier of a SISO system without interference and let P denote the received power, h_r is the channel response at round r and N_0 is the noise power, then

$$I(H_1, \ldots, H_{\mathbb{M}_{max}}) = \sum_{r=1}^{\mathbb{M}_{max}} \log_2 \left(1 + \frac{P|h_r|^2}{N_0}\right). \tag{11}$$

Generalizing the notation from [10], the probability of decoding a transport-block in round r with N_j as the number of dimensions used in round j is

$$\Pr\left(I(H_1, \cdots, H_r) > R_r \sum_{j=1}^{r} N_j, I(H_1, \cdots, H_n) < R_n \sum_{j=1}^{n} N_j, \forall n < r\right) \tag{12}$$

Let $P_{out,n}$ denote the target transport-block error probability after n transmission rounds. The latency constraint is expressed by ensuring that the probability that the transport-block is not served after $\mathbb{M}_{max}$ transmission rounds is below $P_{out,\mathbb{M}_{max}}$. Under this framework, rate adaptation is the optimization of

the rate sequences R_r such that (1) the packet error probability remains below $P_{out,\mathbb{M}_{max}}$ after $\mathbb{M}_{max}$ transmission rounds and (2) the spectral-efficiency is maximized. The optimization is carried out as a function of the distribution of $I\left(H_1,\ldots,H_{\mathbb{M}_{max}}\right)$.

For simplicity, we consider at most two retransmission rounds (ARQ rounds), but our policy can also be applied for more than two. We consider three scenarios

1) Minimal-latency: a trivial case of serving the packet in one round which corresponds to the minimal-latency rate adaptation policy.
2) Latency-constrained with no prior CQI: we consider two transmission rounds and no information about the channel.
3) Latency-constrained with outdated CQI: we consider again two transmission rounds, but unlike the previous case, we assume that we have outdated information about the channel with some correlation with the actual channel.

For simplicity and in the interest of obtaining semi-analytical results, we concentrate on one subcarrier, i.e. that H_r is a scalar.

1) Scenario Analysis: Minimal-latency: We consider first the trivial case of serving the transport-block in one round. This is the minimal-latency rate adaptation policy. The rate allocation law for R_1 is given by the solution to

$$\Pr\left(I(H_1) < R_1\right) = P_{out,1}. \tag{13}$$

Without any a priori information regarding the channel statistics, this essentially says that the best that can be done is to transmit with the lowest spectral-efficiency coding scheme (i.e. lowest MCS) to minimize latency. With a priori information, the largest MCS such that the probability of channel realizations requiring a smaller MCS is still below the threshold is chosen.

Let H_{out} denote the channel corresponding to outdated CQI. If stale CQI is available prior to transmission of the transport-block, then the rate should be chosen such that

$$\Pr\left(I(H_1) < R_1 | H_{out}\right) = P_{out,1} \tag{14}$$

2) Scenario Analysis: Latency-constrained with no Prior CQI: We now consider the case with two transmission rounds. Let B define the number of information bits to be transmitted. Let N_T denote the total number of dimensions available and let N_1 denote the number of dimensions used in the first round.

Hence, the rate in the first round is $R_1 = \frac{\log_2 B}{N_1}$, and the rate in the second round $R_2 = \frac{\log_2 B}{N_T}$. We define $\rho = \frac{N_1}{N_T}$ and we can relate R_1 to R_2 with $R_2 = \rho R_1$.

Let $\bar{R}$ denote the overall spectral efficiency, with $P_{out,1}$ as the outage probability after the first round, we have

$$\bar{R} = R_1 \left(1 - P_{out,1}\right) + P_{out,1} R_2 = R_1 \left(1 - P_{out,1}\right) + P_{out,1} \rho R_1. \tag{15}$$

We want to maximize $\bar{R}$ such that the probability of outage after the second round is below the given constraint $P_{out,2}$. For the first round, there is no feedback information. The outage probability $P_{out,1}$ is unknown but it depends on H_1 and the Signal-to-Noise Ratio (SNR). We can relate R_1 to $P_{out,1}$ as follows. From equation (13), we have

$$\Pr\left(I(H_1) < R_1\right) = \Pr\left(\log_2\left(1 + \mathrm{SNR}|h_1|^2\right) < R_1\right) = P_{out,1}. \tag{16}$$

Consequently, we obtain $R_1 = \log_2\Big(1 - \mathrm{SNR}\ln(1 - P_{out,1})\Big)$.

In the second round, feedback about the previous round is available. The outage probability is now given by

$$\Pr\Big(I(H_1, H_2) < R_2 | I(H_1) < R_1\Big) = P_{out,2} \tag{17}$$

We can rewrite equation (17) as follows

$$\begin{aligned} \Pr\Big(I(H_1, H_2) < R_2 | I(H_1) < R_1\Big) = & \quad \frac{\Pr(I(H_1, H_2) < R_2, I(H_1) < R_1)}{\Pr(I(H_1) < R_1)} \\ = & \quad \frac{\int_0^{\frac{2^{R_1}-1}{\mathrm{SNR}}} e^{-|h_1|^2} d|h_1|^2}{P_{out,1}} - \frac{\int_0^{\frac{2^{R_1}-1}{\mathrm{SNR}}} e^{-a-|h_1|^2} d|h_1|^2}{P_{out,1}} = P_{out,2} \end{aligned} \tag{18}$$

where $a = \left(\left(\frac{2^{R_1}}{1+\mathrm{SNR}|h_1|^2}\right)^{\frac{\rho}{1-\rho}} \frac{1}{\mathrm{SNR}}\right) - \frac{1}{\mathrm{SNR}}$, and the limits stem from the fact that if $I(H_1) < R_1$ then $|h_1|^2 < \frac{2^{R_1}-1}{\mathrm{SNR}}$. The integrals in equation (18) are evaluated numerically.

To find the optimal value of R_1 in the first round, we perform an extensive exploration on $P_{out,1}$, given that we want to maximize equation (15) and subject to the constraint $P_{out,2}$ in equation (18).

3) Scenario Analysis: Latency-constrained with Outdated CQI: Because of sparse traffic characteristics, of moderate to high mobility, of insufficient uplink CQI periodicity or of inter-cell interference, we investigate cases where the UL CQI is outdated or unavailable. In such cases, the scheduler can only benefit from binary feedback after the first HARQ transmission round (in the form of ACK/NACK signaling [2]). For the outdated CQI case, it is assumed that the fading statistics are available to the transmitter. This

assumption is reasonable because the eNB scheduler can maintain a database of channel measurements in its cell, allowing it to derive the fading statistics over time.

We make the additional assumption that the channel remains constant over the two transmission rounds and let $h = h_1 = h_2$. Furthermore, we denote by h_{out} the channel value that corresponds to the outdated CQI. In order to model a possible correlation between h_{out} and h, we use the following model. Let λ be the correlation parameter, then

$$h = \sqrt{\lambda} h_{out} + \sqrt{1-\lambda} h^{'}$$

where h_{out} and $h^{'}$ are i.i.d. Gaussian-distributed random variables. Note that in this case, $\lambda = \mathbb{E}\left[h_{out} h^{*}\right]$. In addition, $|h|^2$ is a non-central Chi-square random variable with two degrees of freedom. We follow the same general procedure to obtain the throughput and probability of outage than in the previous cases. However, the spectral efficiency is a function of the outdated CQI and we have to average over the distribution of $|h_{out}|^2$.

First, let $\gamma_1 = \sqrt{\frac{2^{R_1}-1}{\mathrm{SNR}}}$ be the outage threshold in the first round and $\gamma_2 = \sqrt{\frac{2^{R_2}-1}{\mathrm{SNR}}}$ be the outage threshold in the second round. Then, $\Pr\left(h > \gamma_1\right)$ represents the probability of having a successful transmission in the first round, $\Pr\left(h < \gamma_1, h > \gamma_2\right)$ is the probability of being unsuccessful in the first round but successful in the second round, and $\Pr(h < \gamma_2)$ gives the probability of being in outage. All these probabilities are a function of the Cumulative Distribution Function (CDF) of $|h|^2$. The non-centrality parameter of $|h|^2$ is $s^2 = \lambda |h_{out}|^2$. Let $F_{\chi}\left(\chi\right)$ denote the CDF of $|h|^2$. It can be expressed in terms of the Marcum Q-function [37], i.e.

$$F_{\chi}\left(\chi\right) = 1 - Q_1(s, \chi) \tag{19}$$

where $Q_M(a, b)$ is the Marcum Q-function with M degrees of freedom and paramters a and b.

The overall spectral efficiency $\bar{R}$ over the two ARQ rounds can be written as

$$\bar{R} = \Pr\left(h > \gamma_1\right) R_1 + \Pr\left(h > \gamma_2, h < \gamma_1\right) R_2. \tag{20}$$

To find the optimal rates, we first obtain R_2 from the outage constraint $P_{out,2}$. Since we know that $h < \gamma_2$ implies an outage, R_2 is given by solving equation (21) for R_2. Therefore

$$P_{out,2} = \Pr\left(|h|^2 < \frac{2^{R_2}-1}{\mathrm{SNR}}\right) = F_{\chi}\left(\gamma_2^2\right) \tag{21}$$

Next, to find the value of R_1 that will maximize the overall spectral efficiency, we first write (20) in terms

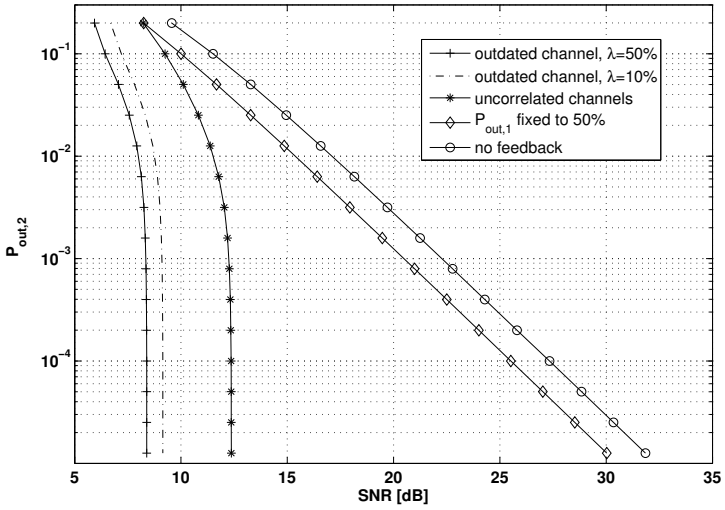


Fig. 3. For different values of the probability of outage after the second round $P_{out,2}$, we calculate the corresponding SNR for the different scenarios. The symbol λ is the correlation coefficient between the actual channel and the channel corresponding to outdated CQI information. We compare a correlation coefficient value of $50\%, 10\%$ and uncorrelated case. For comparison purposes, we also plot two more cases. First when no ACK/NACK feedback is available from the HARQ process. Second, when $P_{out,1}$ is fixed to 50% with $\rho = 0.5$ to make sure that 50% of the dimensions are used in each round.

of the Marcum Q-function. We have

$$\bar{R} = Q_1(a_1, b_1)R_1 + (Q_1(a_2, b_2) - Q_1(a_1, b_1))R_2 \tag{22}$$

where $a_1 = a_2 = s$, $b_1 = \gamma_1$, and $b_2 = \gamma_2$. We now take the derivative of equation (22) with respect to R_1, and we solve for R_1 when the derivative is zero. We obtain

$$\begin{aligned} \frac{\partial \bar{R}}{\partial R_1} &= \frac{\partial Q_1(a_1, b_1)}{\partial R_1} R_1 + Q_1(a_1, b_1) - \frac{\partial Q_1(a_1, b_1)}{\partial R_1} R_2 \\ &= \frac{\partial Q_1(a_1, b_1)}{\partial R_1} (R_2 - R_1) + Q_1(a_1, b_1) = 0. \end{aligned} \tag{23}$$

To find the derivative of the Marcum Q-function in (23), we used [38].

4) Numerical Results: In this section, we present numerical results in terms of (1) the probability of outage and (2) the achieved spectral efficiency. Remember that we assume Gaussian signaling to compute the mutual information. Throughout this section, we fix the spectral efficiency to 2 bits per channel use. The maximum number of retransmissions is one round (at most two transmission rounds).

Figure 3 presents the minimum SNR necessary to achieve a given outage probability $P_{out,2}$. For a given value of $P_{out,2}$, we calculate the corresponding SNR for our rate adaptation policy. For comparison purposes, we consider two more cases. First we evaluate a case where we force the probability of outage after the first round to 50%, fixing $\rho = 0.5$ to make sure that 50% of the dimensions are used in each

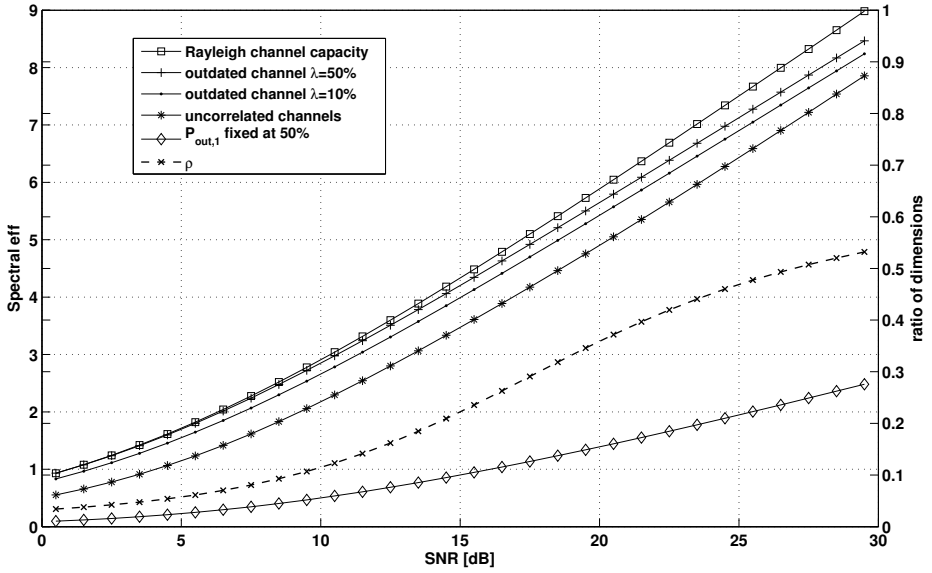


Fig. 4. The axis on the left (solid lines) shows the spectral efficiency versus SNR for the different scenarios. We set $P_{out,2}$ to 1%. The symbol λ is the correlation coefficient between the actual channel and the outdated/stale CQI information. We compare correlation coefficients of $50\%, 10\%$ and uncorrelated case. For comparison purposes we plot the curve for the ergodic capacity and $P_{out,1}$ fixed to 50% with $\rho = 0.5$ to make sure that 50% of the dimensions are used in each round. The axis on the right (dashed lines) shows the ratio of dimensions used in the two rounds.

round. Typically, while conventional systems try to ensure a 10% outage probability per slot, we observe from our results that a higher value gives, in fact, a higher overall spectral efficiency. Second, we evaluate a case where no feedback at all is available i.e. when we can not even receive ACK/NACK from the HARQ process. This highlights the significant gain from adapting the rate across rounds with only one bit of feedback, even in the case without any CQI information. The gain is even higher when only outdated CQI information is available. Our rate adaptation policy gives a zero probability of outage without the need of having a high SNR. From the results in figure 3, we can observe that it does not make a difference to increase the SNR above 12.5 dB for the case without CQI. We show that adjusting the dimensions used in each round results in almost causal feedback performance. In our scenarios, the two rates are simply controlled by $\rho = \frac{R_2}{R_1}$, which depends on the SNR and target outage probability $P_{out,2}$. We only need one bit of feedback, which we get causally from HARQ. In fact, it is the state of the channel that chooses the code rate. By choosing the rate in the first round as high as possible, we can guarantee a probability of outage after the second round while maximizing the spectral efficiency.

Figure 4 presents the overall spectral efficiency obtained for a given SNR. We set $P_{out,2}$ to 1%. For the outdated CQI case, we consider $\lambda = 50\%$ and $\lambda = 10\%$. For reference purposes, we also plot the ergodic capacity (Rayleigh channel capacity), i.e. perfect rate adaptation. Finally, we consider a scenario where

the rate in the first round is chosen as the one that corresponds to a probability of outage after the first round of 50%. This value is chosen because it gives the highest spectral efficiency. Fixing the probability of outage after the first round to more or less than 50% gives, in fact, a lower overall spectral efficiency.

From our results, we see a significant improvement in spectral efficiency even in the case without CQI. When we can benefit from outdated CQI, we achieve a performance close to the ergodic capacity. If we look at the results for the ratio of dimensions for the second round, we can see that as the SNR gets higher, more dimensions are used in the second round.

B. Interference networks analysis

We now consider one interferer and we model it with an activity factor, which means that the interferer could be active or inactive. The activity factor is Bernoulli distributed with probability p. The rate with Gaussian codebooks that can be achieved by the protocol depends on the interference state (interference active or inactive). Let R_H be the capacity that can be achieved without interference, and R_L the corresponding capacity with interference, which are given by:

$$R_H = \log_2(1 + \text{SNR}_1) \tag{24}$$

$$R_L = \log_2\left(1 + \frac{\text{SNR}_1}{1 + \text{SNR}_2}\right) \tag{25}$$

where SNR_1 is the SNR for the user of interest, SNR_2 is the corresponding SNR for the interferer and we assume unitary noise variance. We consider a HARQ protocol with two rounds and we define $\rho = \frac{N_1}{N_T}$, where N_1 is the number of dimensions used in the first round and N_T has been previously defined as the total number of dimensions. Then for a packet of length B bits, the rate in the first round is $R_1 = \frac{1}{\rho N_T}\log_2 B = R_H$ and in the second round $R_2 = \frac{1}{N_T}\log_2 B = R_L$. Therefore, $R_L = \rho R_H$, and $\rho = \frac{R_L}{R_H}$.

In sections V-B1 and V-B2, we derive the throughput assuming there is no outage at the end of the HARQ protocol with and without feedback and in section V-B3, we give the expressions for the case of a residual outage at the end of the protocol with feedback.

1) Zero-outage throughput without feedback and no delay constraint: We now look at the case of no feedback. Let the activity factor μ define the state of the interference. We consider ON/OFF interference, therefore, if $\mu = 0$ there is no interference and $\mu = 1$ means interference is active and this happens with

probability p. Then the throughput $\bar{R}$ with zero outage (without delay) is given by:

$$\bar{R} = \mathbb{E}_{\mu} I(X;Y|\mu) = (1-p)R_H + pR_L \tag{26}$$

It is interesting to note that (26) is the ergodic capacity (average over all possible states). In the next section, we explore the case when feedback becomes available and we look at the case of more than two transmission rounds.

2) Zero-outage throughput with feedback: In this case, we assume that we have feedback from the HARQ protocol and we vary the tolerable latency by fixing the maximum number of transmission rounds $\mathbb{M}_{max}$, but still assume zero-outage probability. Then, given that we want zero-outage at round $\mathbb{M}_{max}$, we choose the rate that guarantees successful decoding (i.e. R_L). We choose the rate in the first round to be as high as possible, and the intermediate rates are at the optimal value between R_L and R_H. Therefore, the rate after the r^{th} round is given by:

$$R_1 = \frac{\log_2 B}{\rho_1 N_T} = R_H \qquad r = 1 \tag{27}$$

$$R_r = \frac{\log_2 B}{(\sum_{j=1}^{r} \rho_j) N_T} = \left(\frac{\rho_1}{\sum_{j=1}^{r} \rho_j} \right) R_H \qquad 2 \leq r < \mathbb{M}_{max} \tag{28}$$

$$R_{\mathbb{M}_{max}} = \frac{\log_2 B}{N_T} = R_L = \rho_1 R_H \;\Rightarrow \rho_1 = \frac{R_L}{R_H} \qquad r = \mathbb{M}_{max} \tag{29}$$

In this case, the throughput expression for $\mathbb{M}_{max}$ rounds is given by:

$$\bar{R} = (1-p)R_H + \sum_{r=2}^{\mathbb{M}_{max}-1} p^{r-1}(1-p) \left(\frac{\rho_1}{\sum_{j=1}^{r} \rho_j} \right) R_H + p^{(\mathbb{M}_{max}-1)} R_L \tag{30}$$

For the rates to be achievable, we observe that there is a restriction on the ratio of dimensions after the second round ρ_r, $r > 1$. This restriction comes from the fact that the rate after round r is $\left(\sum_{j=1}^{r-1} \rho_j \right) N_T R_L + \rho_r N_T R_H$ which means that after round r we decode if:

$$\left(\sum_{j=1}^{r} \rho_j \right) N_T R_r < \left(\sum_{j=1}^{r-1} \rho_j \right) N_T R_L + \rho_r N_T R_H \tag{31}$$

$$R_r = \left(\frac{\rho_1}{\sum_{j=1}^{r} \rho_j} \right) R_H < \frac{\left(\sum_{j=1}^{r-1} \rho_j \right) R_L + \rho_r R_H}{\sum_{j=1}^{r} \rho_j}$$

$$\rho_r > \rho_1 \left(1 - \sum_{j=1}^{r} \rho_j \right) \tag{32}$$

If we look at figure 5, the solid lines and the right axis show the zero-outage throughput for the HARQ protocol with a maximum number of rounds $\mathbb{M}_{max} = \{1, 2, 3, 4\}$. We can see that there is a high gain when going from one to two rounds and after three rounds there is only a marginal gain. For reference purposes, we also plot the average rate that can be achieved when the probability of interference is 50% and we can see that after 2 rounds, we can obtain a higher spectral efficiency. The dashed lines and left axis show how the dimensions are being distributed across the rounds of the protocol. We illustrate the case of three rounds (i.e. $\mathbb{M}_{max} = 3$) and we look at the proportion of physical dimensions used in each round (ρ_r). In both cases, the interference strength is the same as the user of interest $(\text{SNR}_1 = \text{SNR}_2)$, the channel is AWGN and we assume Gaussian signals with one interferer active with probability $p = 0.5$. If we look at 10 dB SNR, we observe that 20% of the dimensions are used in the first round, 15% in the second round and the remaining 65% are left for the third round. In general, we observe that at high SNR, the number of dimensions used in the first round decreases, leaving progressively more dimensions to the last round. If we think about the almost blank subframes (ABS) feature of LTE, which restricts the transmission in the cell if there is interference, we would have a lower spectral efficiency than the average rate and therefore, by adapting the dimensions one can achieve a higher spectral efficiency even in the presence of interference.

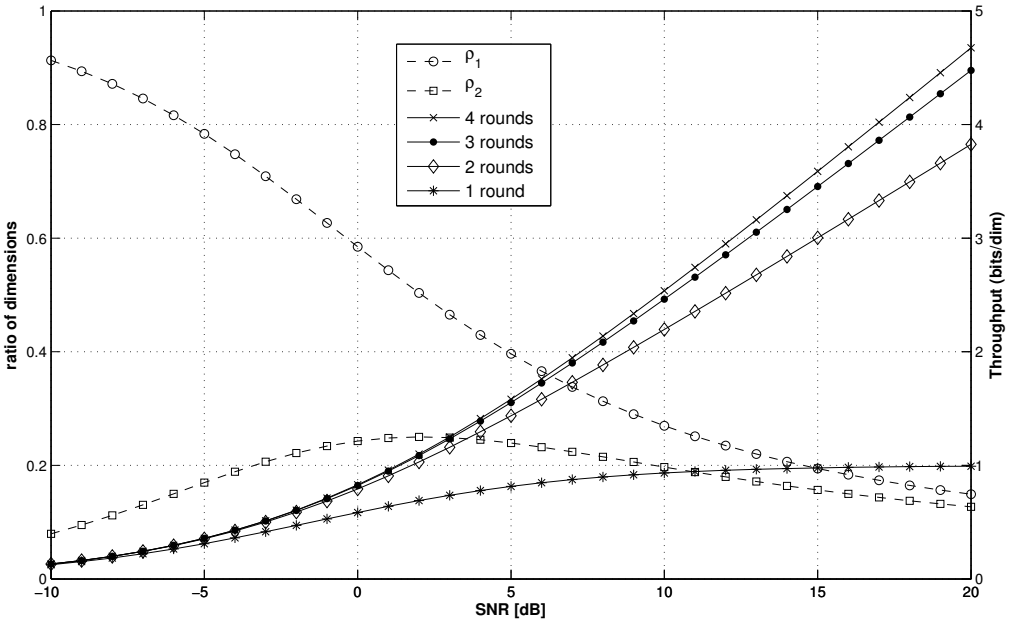


Fig. 5. The axis on the right (solid lines) shows the zero-outage throughput for the HARQ protocol with different number of rounds, while the axis on the left (dashed lines) shows the ratio of dimensions per round for the three rounds, zero-outage HARQ protocol. In both cases the channel is AWGN with Gaussian signals and there is one interferer with probability $p = 0.5$. The interference strength is the same as the user of interest $(\text{SNR}_1 = \text{SNR}_2)$.

3) Throughput with outage and feedback: In this case, we allow the protocol to have a residual outage probability which is overcome by an upper layer ARQ process on top of the IR-HARQ [34], [10], and we assume that we have feedback. For two rounds, from (8) the throughput is given by:

$$\bar{R} = (1 - P_{out}(\rho, R_2)) \left[(1 - P_{out,1}(\rho, R_2)) \frac{R_2}{\rho} + P_{out,1} \left(1 - P_{out,2}(\rho, R_2|out_1)\right) R_2 \right] \tag{33}$$

where $P_{out,2}(\rho, R_2|out_1)$ is the outage probability at the second round, given that there was an outage in the first round, and $P_{out,r}$ is the probability of outage at round r.

$I(\mu_r)$ is the mutual information as a function of the state of the interference at round r, and it is defined by μ_r:

$$I(\mu_r) = \begin{cases} \log_2(1 + \text{SNR}_1) & \mu_r = 0 \\ \log_2\left(1 + \frac{\text{SNR}_1}{1+\text{SNR}_2}\right) & \mu_r = 1 \end{cases} \tag{34}$$

Now, we can define the probabilities in (33) where $P_{out,1}(\rho, R_2)$ is the outage probability at the first round and it is given by:

$$P_{out,1}(\rho, R_2) = \Pr(R_2 > \rho I(\mu_1)) = \begin{cases} 1 & \text{if } R_2 > \rho I(0) \\ 0 & \text{if } R_2 < \rho I(1) \\ \Pr(\mu_1 = 1) = p & \text{if } \rho I(1) < R_2 < \rho I(0) \end{cases}$$

$P_{out,2}(\rho, R_2|out_1)$, the outage probability at the second round given that there was an outage in the first round, is given by:

$$\begin{aligned} P_{out,2}(\rho, R_2|out_1) &= \Pr(R_2 > \rho I(\mu_1) + (1-\rho) I(\mu_2) | R_2 > \rho I(\mu_1)) \\ &= \begin{cases} 1 & R_2 < \rho I(1) \\ \Pr((1-\rho) I(\mu_2) < R_2 - \rho I(1)) & \rho I(1) < R_2 < \rho I(0) \end{cases} \end{aligned}$$

$$\text{where } \Pr((1-\rho) I(\mu_2) + \rho I(1) < R_2) = \begin{cases} p & \text{if } I(1) < R_2 < \rho I(1) + (1-\rho) I(0) \\ 0 & \text{if } I(1) > R_2 \\ 1 & R_2 > \rho I(1) + (1-\rho) I(0) \end{cases}$$

Finally, $P_{out}(\rho, R_2)$ is the probability of outage after the second round, independently of the interference state at the first round and it is given by:

$$P_{out}(\rho, R_2) = \begin{cases} p(1-p) & R_2 > \rho I(1) + (1-\rho) I(0) \\ p^2 & R_2 < \rho I(1) \end{cases} \tag{35}$$

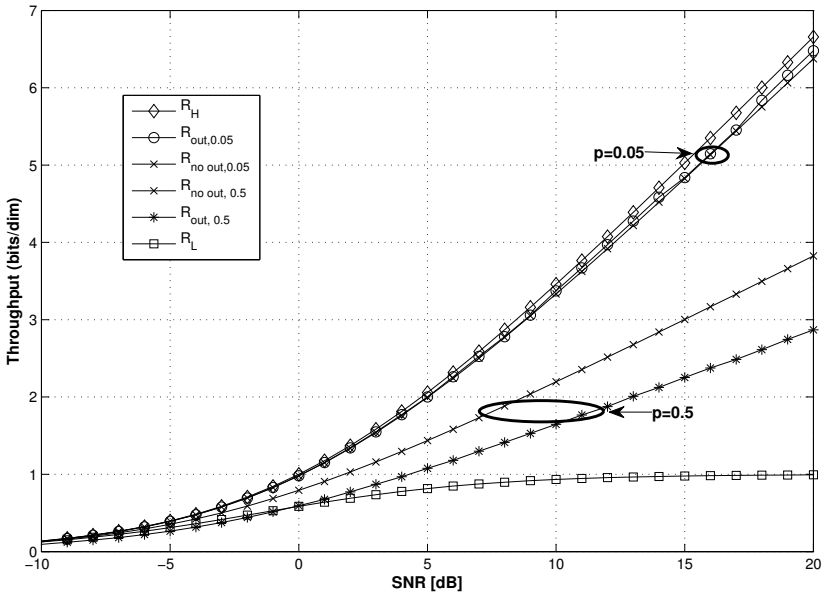


Fig. 6. Throughput of the two rounds HARQ protocol in an AWGN channel with Gaussian signals. There is one interferer with probability $p = 0.05, 0.5$.

Figure 6 shows the throughput of the HARQ protocol with two transmission rounds. There is one interferer with probability $p = \{0.05, 0.5\}$. We compare the zero-outage throughput against the throughput that allows an outage at the end of the protocol. We also plot the maximum capacity achieved with one round and no interference R_H and the corresponding capacity for interference R_L. If we look at the case of 50% probability of interference, we can see that the zero-outage throughput is higher for all SNR values, however, if we look at the case with a lower probability of having interference ($p = 0.05$, or 5%), we have almost the same throughput, except at high SNR, where the throughput with an outage is slightly higher. In this case, we also see that the capacity that can be achieved by adapting the rate and dimensions gets close to the capacity achieved without interference.

4) Discussion: If we consider the case with the ergodic capacity and no feedback, we transmit N_T dimensions per channel realization. Therefore, we have the average capacity:

$$\mathbb{E}_\mu = \begin{cases} \log_2(1 + \text{SNR}_1) & \mu = 0 \\ \log_2\left(1 + \frac{\text{SNR}_1}{1+\text{SNR}_2}\right) & \mu = 1 \end{cases} \tag{36}$$

where μ is the state of the interference. Now, if we consider a channel with feedback of the state of the interference (non-causal feedback). Then at round r, the transmit signal is a function of the message W

and the interference state μ:

$$\begin{cases} x_r = f(W, \mu) & r > 1 \\ x_1 = f(W) \end{cases} \tag{37}$$

To get an insight into how a rate-adaptive scheme performs when changing the number of dimensions across rounds, we focus on the case of the HARQ protocol with two transmission rounds. At round r, if $\mu = 1$, then there is no transmission, and it happens with probability $\Pr(\mu_r = 1) = p$. However, if there is no interference, $\mu_r = 0$, it transmits with $\frac{N_T}{1-p}$ dimensions, and in this case we get a throughput$=$ $(1-p)\left(\frac{\log_2(1+\text{SNR}_1)\frac{N_T}{(1-p)}}{N_T}\right) = \log_2(1+\text{SNR}_1)$ which is the maximum achievable spectral efficiency. When feedback becomes available, it allows the scheme to perform better than the ergodic capacity. The latter is in contrast to the work in [10] where in the infinite delay case, the authors conclude that the maximum that can be achieved is the ergodic capacity. The difference comes from the fact that in [10] there is always a fixed bandwidth allocation for each user, regardless of the state of the channel. In our case, we dynamically adapt the bandwidth for each user depending on the interference conditions of past transmissions for the same codeword. From the perspective of the scheduler, the bandwidth is better distributed. From our initial analysis with interference, we can conclude that the highest spectral efficiency that can be achieved happens in the case of the zero-outage protocol where increasing the delay becomes beneficial to a certain point and brings only a marginal gain after this point. In the next section, we look at practical interference scenarios where having zero-outage throughput is not possible since power control and channel state feedback are not assumed [16]. However, a constraint on the outage probability can be imposed. We also look at the difference between the UL and DL channels.

VI. Rate optimization in practical scenarios

To model a practical setting, we consider signals coming from discrete alphabets. Indeed, while Gaussian signals achieve the maximum spectral efficiency, deployed systems such as LTE or HSPA make use of small, finite-size input alphabets. We look at the case of two transmissions rounds where for a given R_1, we have successful transmission in the first round if $\Pr(I'_1(H_1) > R_1)$, after the second round, outage corresponds to $\Pr(I'_2(H_2) < R_2)$. Let $R_1 = \frac{\log_2 B}{L_1 T_{dim}}$ be the rate at the first round, and $R_2 = \frac{\log_2 B}{N_T T_{dim}}$ the rate at the second round. Then, the overall throughput expression is:

$$\bar{R} = \left(\Pr(I'_1(H_1)) > R_1\right) R_1 + \left(\Pr(I'_1(H_1) < R_1, I'_2(H_2) > R_2\right) R_2 \tag{38}$$

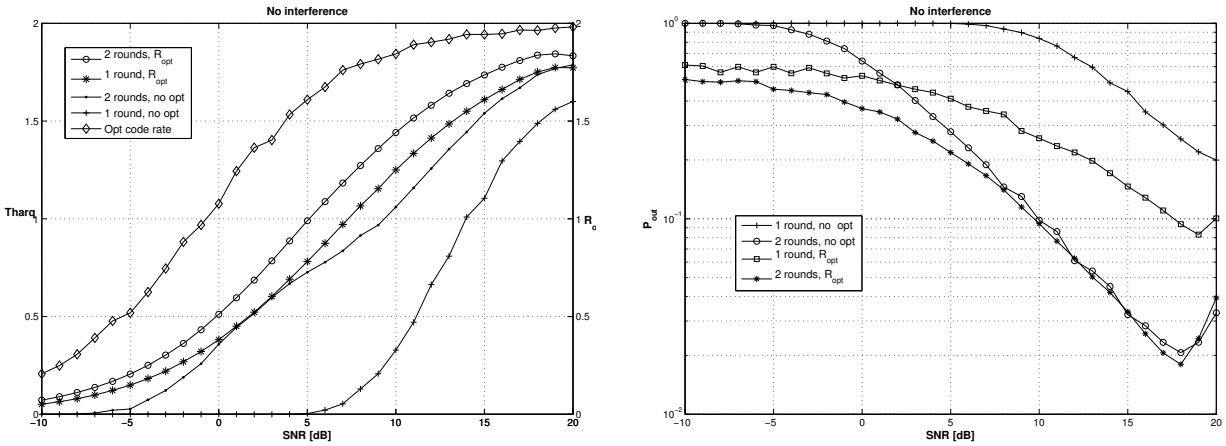
where the outage probability is $P_{out} = \Pr(I_1'(H_1) < R_1, I_2'(H_2) < R_2)$.

For the sake of obtaining long-term average throughput, we isolate the target channel h and average (38) over the channel distribution: $\Pr\left(I_1'(H_1) > R_1\right) = \mathbb{E}_H \Pr\left(I_1'(H_1) > R_1 | h\right)$.

We start with the case of fixed rates across rounds. In this case the throughput is given by the special case of (38) when $R_1 = R_2$.

A. Rate optimization (fixed across rounds)

We can now proceed to optimize the rate for different operating SNR points. This means that at every SNR point, we choose the rate that gives the maximum throughput.



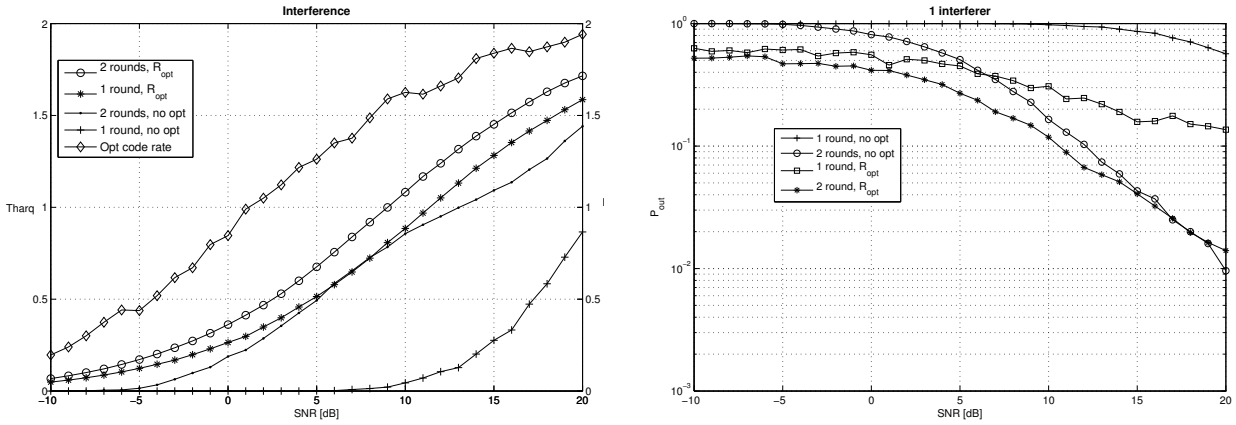
(a) Rate optimization, no interference.

(b) Probability of outage, no interference

Fig. 7. In (a) we show the rate optimization of the HARQ protocol for a diferent number of rounds $\mathbb{M}_{max} = 1, 2$ in a Rayleigh fading channel with QPSK modulation. The rates are fixed across rounds $R_1 = R_2 = R$. Figure (b) shows the corresponding probability of outage.

In figure 7, we show the spectral efficiency and the probability of outage when we optimize the rate per transmission. The rate is the same across rounds. We compare rate optimization with a fixed rate operation for a maximum number of HARQ rounds $\mathbb{M}_{max} = 1, 2$ and we can see that, for example, optimizing the rate with one HARQ round gives more or less the same gain as having an additional transmission round, but minimizing the delay. If we allow two HARQ rounds with rate optimization, then the gain in throughput is even higher. In this case, the channel is constant and assumed unknown to the transmitter.

Figure 8 shows the spectral efficiency and the probability of outage with one or two transmission rounds in a Rayleigh fading channel with QPSK modulation. Across the transmission rounds, the rates are fixed. We consider that there is one interferer present all the time, so the activity factor $\mu = 1$. If we allow the activity factors to take other values in time than one, then we can talk about the DL channel.



(a) Rate optimization, one interferer.

(b) Probability of outage, one interferer

Fig. 8. In (a) we show the rate optimization of the HARQ protocol for a diferent number of rounds $\mathbb{M}_{max} = 1, 2$ in a Rayleigh fading channel with QPSK modulation. The rates are fixed across rounds $R_1 = R_2 = R$. There is one interferer all the time. Figure (b) shows the corresponding probability of outage.

In figure 9 we show the comparison between rate optimization for constant and iid channels across the HARQ rounds. We assume that there is no outage constraint, and we choose the rates that will maximize the throughput independently of the residual outage probability at the end of the protocol. Although the gain is sligthly higher for iid channels (around 3 dB), there is always a gain in throughput for the whole SNR range. We observe that the gain is higher in the high SNR region. As a next step, we proceed to

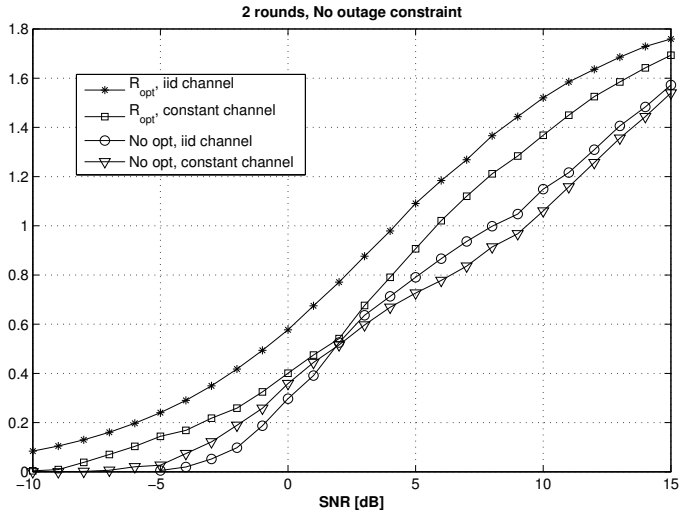


Fig. 9. Rate optimization for both constant channel and iid channel across the HARQ rounds.

optimize the number of dimensions used in each of the HARQ rounds.

B. Rate optimization with an outage constraint

In some LTE scenarios, requirements on the latency can be imposed which can be inmediately translated into the outage at the end of the retransmission protocol. Depending on the type of application, traffic, etc, the latency requirements will be different. Therefore, we study two different scenarios, one with a relatively relaxed outage constraint of 10% and one with a more strict constraint of 1%. To model this constraint, we consider a retransmission protocol with a maximum of $\mathbb{M}_{max}$ rounds, and we say that the constraint is met whenever the packet error probability after $\mathbb{M}_{max}$ rounds is smaller than a predefined threshold P_{out}. To find the optimal rates, we look at the case of a two rounds HARQ protocol. We start by choosing the rate in the second round as the rate that satisfies the outage constraint, i.e. we solve $\Pr(I'_{2,i}(H) < R_2)$ for R_2. The rate in the first round, R_1, is chosen as the one that maximizes the throughput expression in (38) while satisfying the given constraint. In this case, we also optimize the number of dimensions used in each of the retransmission rounds. Since there is no closed-form expression for the probability of outage for discrete signals, we notice that $\Pr(I(H_2) < R_2)$ represents the CDF of the mutual information evaluated at R_2, $F_I(R_2)$. With the help of the inversion formula in [39], we use the characteristic function of the mutual information $\Phi_I(\omega)$ to find the CDF as:

$$F_I(R_2) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\Im\left\{\exp(-j\omega R_2)\Phi_I(\omega)\right\}}{\omega} d\omega \tag{39}$$

The characteristic function is defined as $\Phi_I(\omega) = \mathbb{E}\left[\exp(j\omega I)\right]$, where $\mathbb{E}$ denotes expectation. Since we assume Rayleigh fading, the expectation is over the channel squared magnitude probability density function (PDF), which is exponentially distributed:

$$\Phi_I(\omega) = \int_0^\infty \exp(j\omega I(h))\exp(-h)dh \tag{40}$$

We now give an example on how to obtain the rates for the case of iid channels across the HARQ rounds. According to the rate definition in (1), we define the rates R_1 and R_2 as $R_1 = \frac{\log_2 B}{T_{\dim} L_1}$ and $R_2 = \frac{\log_2 B}{T_{\dim}(L_1+L_2)}$ where L_r is the number of dimensions used in round r. If we define $\rho = \frac{L_1}{L_1+L_2}$ (i.e. $R_2 = \rho R_1$), then, the mutual information in bits/dim per round is given by:

$$I'_1(H_1) = \frac{T_{\dim} L_1 I_1(H_1)}{T_{\dim} L_1} = I_1(H_1) \tag{41}$$

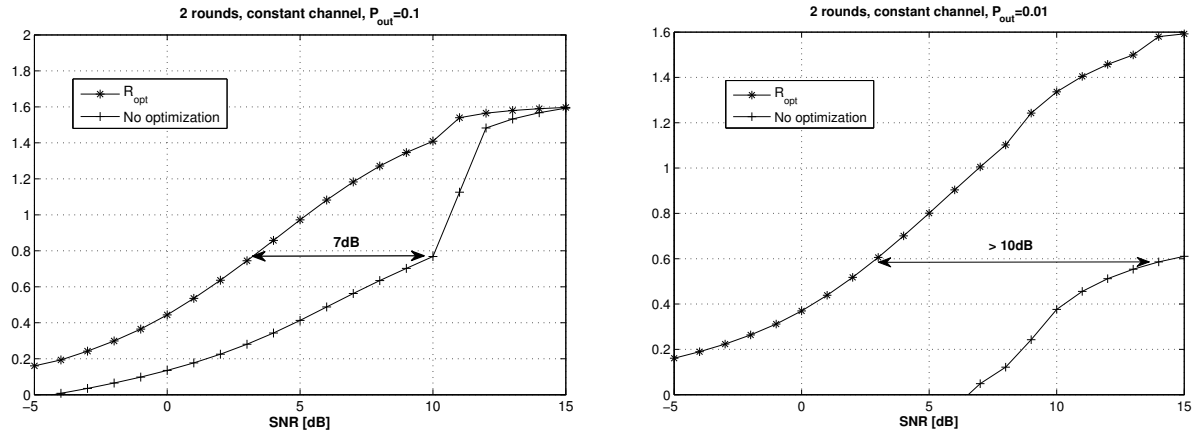
$$I'_2(H_2) = \frac{T_{\dim} \times (L_1 I_1(H_1) + L_2 I_2(H_2))}{T_{\dim}(L_1+L_2)} = I_1(H_1) + I_2(H_2)\left(\frac{1-\rho}{\rho}\right) \tag{42}$$

In this case, the characteristic function of the mutual information at the second round is given by:

$$\phi_{I_2}(\omega) = E\left[\exp(j\omega I_2'(H_2))\right] \tag{43}$$
$$= E\left[\exp\left(j\omega\left(I_1(H_1) + I_2(H_2)\left(\frac{1-\rho}{\rho}\right)\right)\right)\right]$$
$$= E\left[\exp\left(j\omega I_1(H_1)\right)\exp\left(j\omega I_2(H_2)\left(\frac{1-\rho}{\rho}\right)\right)\right] \tag{44}$$
$$= E\left[\exp\left(j\omega I_1(H_1)\right)\right] E\left[\exp\left(j\omega I_2(H_2)\right)^{\left(\frac{1-\rho}{\rho}\right)}\right] \tag{45}$$
$$= \phi_{I_1}(\omega)\phi_{I_2}\left(\omega\left(\frac{1-\rho}{\rho}\right)\right)$$

where going from (44) to (45) is possible because the channels are independent across rounds. Finally, we can use (39) and the outage constraint to find R_2.

In figure 10, we show the results for the rate optimization with an outage constraint of 10% and 1% when the channel remains constant across the HARQ rounds, and in figure 11, we show the results for iid channels. There is a clear advantage on optimizing the rate across the HARQ rounds with a maximum gain of more than 10 dB for the constant channel case. The gain is higher when we have a more strict outage constraint of 1%. If we compare these results to those in section VI-A, without outage constraint, we can see the significant gain introduced by optimizing the rate for latency-constrained scenarios. It can also be noticed that the throughput with rate optimization and an outage constraint of 1% is only lower by a small quantity as compared to the outage constraint of 10%. This last observation tells us that optimizing the rate can, indirectly, minimize the latency constraint (achieving almost the same throughput for both outage constraints). In the case of iid channels, we observe a slightly smaller gain with an outage



(a) Rate optimization with an Outage constraint of 10%.

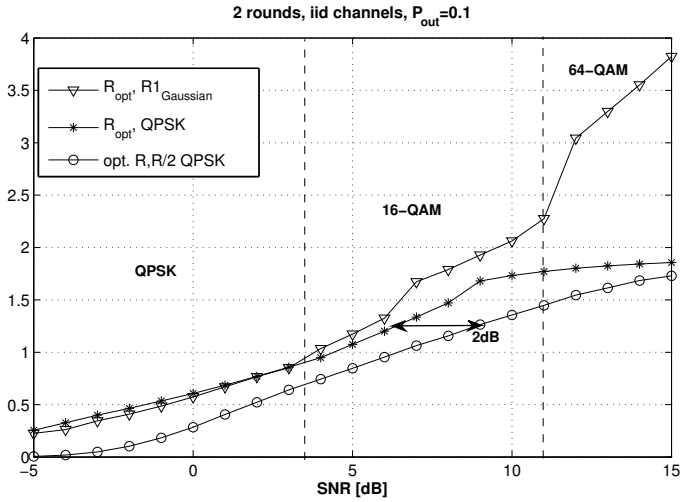
(b) Rate optimization with an Outage constraint of 1%.

Fig. 10. In (a) we show the rate optimization with an outage constraint of 10%. The channel is constant across the HARQ rounds and there is no interference. This is equivalent to a non-line-of-sight (NLOS) with slow fading channel. Figure (b) shows the corresponding curves for an outage constraint of 1%

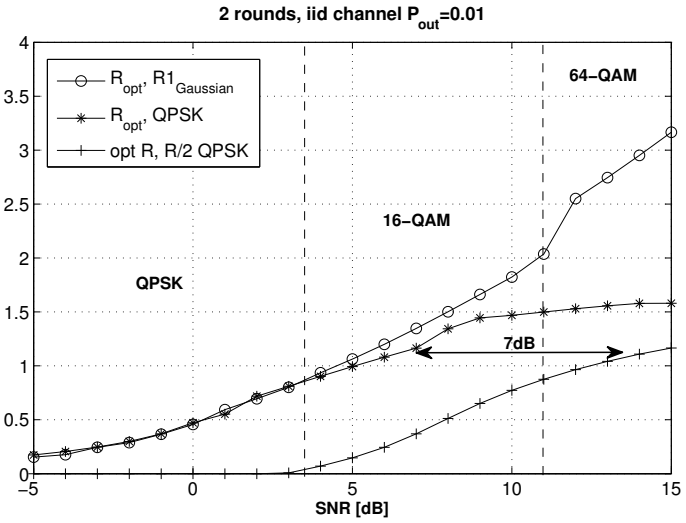
constraint of 10%, so the optimization has a more significant impact for the case of 1%.

In figure 11, to obtain the maximum throughput, we remove the constraint on the modulation. To do

this, we choose the rate in the first round according to the mutual information expression for Gaussian inputs, which is not bounded by a particular modulation order, and we choose the modulation that allows us to achieve this rate. We define threshold values for changing modulation between QPSK, 16-QAM and 64-QAM according to the maximum rate achieved with each modulation for a particular SNR value. We confirm once more that the gain is higher for the 1% case (around 7 dB for the case of a more strict



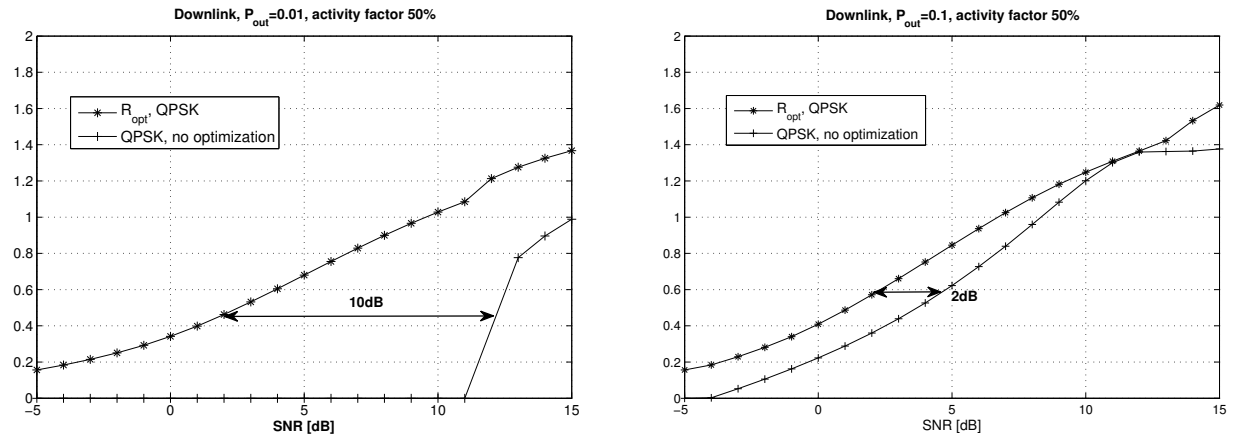
(a) Rate optimization with an Outage constraint of 10%.



(b) Rate optimization with an Outage constraint of 1%.

Fig. 11. In (a) we show the rate optimization with an outage constraint of 10%. The channel is iid across the HARQ rounds and there is no interference. In this case, it is equivalent to a NLOS channel with fast fading or frequency hopping. Figure (b) shows the corresponding curves for an outage constraint of 1%

outage constraint of 1% while a gain of around 2 dB is observed for the 10% case). When we change the modulation with respect to the SNR, we observe a higher throughput in the high SNR region. This is caused by allowing the protocol to use higher modulation orders. In the following section, we show results when considering interference in the scenario.



(a) outage constraint=1% (b) outage constraint=10%

Fig. 12. In (a), we show the rate optimization for Rayleigh fading on the downlink channel with an outage constraint of 1%. And in (b), we show the corresponding results for an outage constraint of 10%. In both cases there is one dominant interferer with an activity factor of 50%.

C. Interference case

We consider one dominant interferer and we assume a constant channel on the desired user. We analyze the case for the UL and DL differently. We first consider the DL of a femtocell with one interferer, and we consider an activity factor for the interferer which means that it is active only a portion of the time. Figure 12(a) shows the results for an activity factor of 50%, i.e. interference is present only half of the time. We fix the outage constraint at 1% and we plot the throughput for the user of interest. In figure 12(b), we show the DL case for an outage constraint of 10%. We observe the same behavior as with the no interference case, with a higher gain for a more strict outage constraint (2 dB for the 10% case against 10 dB for the 1% case).

For the UL, since the interference is coming from other users, it changes in time. Therefore, we consider the interference iid across the HARQ rounds. Figure 13 shows the results for rate optimization in the 10% outage constraint case.

VII. Conclusions

We developed a mathematical framework for the analysis of HetNets and we motivated the use of inter-round resource allocation. Under this framework, we derived distributed resource allocation policies that are applicable for the UL and DL channels. Our policies are based on the dynamic adaptation of the physical dimensions and coding rate used in each round of the IR-HARQ protocol. We considered interference that can be bursty due to the characteristics of HetNets deployments. Rather than performing extensive simulations, we derived analytical expressions, based on mutual information modeling, that capture the

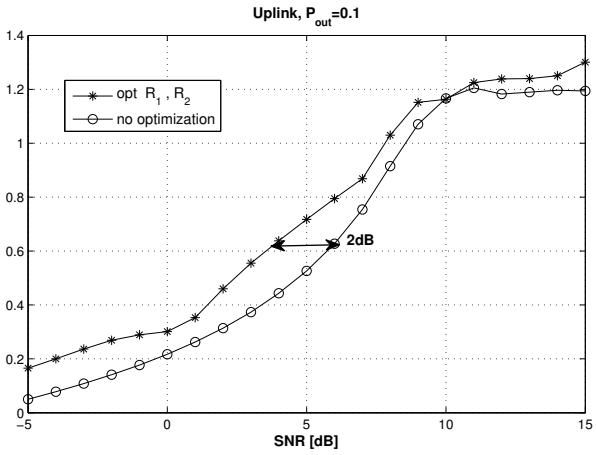


Fig. 13. Rate optimization for Rayleigh fading on the uplink channel. There is one dominant iid interferer. The outage constraint is 10%

throughput performance of HetNets with or without a latency constraint. This mutual information modeling can also be used to derive physical layer abstraction models which can speed up large-scale system simulations that can be extremely time consuming and in some cases computationally unfeasible. We are currently investigating the performance of our resource allocation strategies on practical LTE MODEMs in an interference environment under the constraints of LTE coded-modulation [31], [40].

REFERENCES

[1] S. Lin, D. Costello, and M. Miller, "Automatic-repeat-request error-control schemes," *Communications Magazine, IEEE*, vol. 22, no. 12, pp. 5 –17, december 1984.

[2] S. Sesia, I. Toufik, and M. Baker, *LTE, The UMTS Long Term Evolution: From Theory to Practice*. Wiley Publishing, 2009.

[3] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, *3G Evolution HSPA and LTE for Mobile Broadband*, 1st ed. Academic Press, 2007.

[4] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews, P. Xia, H.-S. Jo, H. S. Dhillon, and T. D. Novlan, "Heterogeneous cellular networks: From theory to practice." *IEEE Communications Magazine*, vol. 50, no. 6, pp. 54–64, 2012.

[5] L. Lindbom, R. Love, S. Krishnamurthy, C. Yao, N. Miki, and V. Chandrasekhar, "Enhanced inter-cell interference coordination for heterogeneous networks in lte-advanced: A survey," *CoRR*, vol. abs/1112.1344, 2011.

[6] H. Arslan, *Cognitive Radio, Software Defined Radio, and Adaptive Wireless Systems (Signals and Communication Technology)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[7] G. P. Koudouridis and H. Li, "Distributed power on-off optimisation for heterogeneous networks - a comparison of autonomous and cooperative optimisation." in *CAMAD*. IEEE, 2012, pp. 312–317.

[8] A. Goldsmith, "The capacity of downlink fading channels with variable rate and power," *Vehicular Technology, IEEE Transactions on*, vol. 46, no. 3, pp. 569 –580, aug 1997.

[9] E. Ohlmer and G. Fettweis, "Rate adaptation for interference cancelation receivers in slowly time-variant mimo channels," in *Sarnoff Symposium (SARNOFF), 2012 35th IEEE*, may 2012, pp. 1 –5.

[10] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *Information Theory, IEEE Transactions on*, vol. 47, no. 5, pp. 1971 –1988, july 2001.

[11] D. Tuninetti, "Transmitter channel state information and repetition protocols in block fading channels," in *Information Theory Workshop, 2007. ITW '07. IEEE*, sept. 2007, pp. 505 –510.

[12] P. Wu and N. Jindal, "Performance of hybrid-arq in block-fading channels: A fixed outage probability analysis," *Communications, IEEE Transactions on*, vol. 58, no. 4, April 2010.

[13] N. Gopalakrishnan and S. Gelfand, "Rate selection algorithms for ir hybrid arq," in *Sarnoff Symposium, 2008 IEEE*, april 2008, pp. 1 –6.

[14] S. M. Kim, W. Choi, T. W. Ban, and D. K. Sung, "Optimal rate adaptation for hybrid arq in time-correlated rayleigh fading channels," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 3, pp. 968 –979, march 2011.

[15] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theor.*, vol. 48, no. 5, pp. 1135–1149, sep 2006. [Online]. Available: http://dx.doi.org/10.1109/18.995554

[16] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *Information Theory, IEEE Transactions on*, vol. 45, no. 5, pp. 1468 –1489, jul 1999.

[17] I. Bettesh and S. Shamai, "Optimal power and rate control for minimal average delay: The single-user case," *Information Theory, IEEE Transactions on*, vol. 52, no. 9, pp. 4115 –4141, sept. 2006.

[18] C. Chai, T. T. Tjhung, and L. C. Leck, "Combined power and rate adaptation for wireless cellular systems," *Wireless Communications, IEEE Transactions on*, vol. 4, no. 1, pp. 6 – 13, jan. 2005.

[19] T. Chaitanya and E. Larsson, "Outage-optimal power allocation for hybrid arq with incremental redundancy," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 7, pp. 2069 –2074, july 2011.

[20] L. Szczecinski, C. Correa, and L. Ahumada, "Variable-rate retransmissions for incremental redundancy hybrid arq," *CoRR*, vol. abs/1207.0229, 2012.

[21] *D4.3 Adaptive Modulation and Coding Scheme and Hybrid ARQ Mechanism*, EU FP7 Project LOLA (Achieving Low-Latency in Wireless Communications), v3.0, January 2013.

[22] U. Erez, M. Trott, and G. W. Wornell, "Rateless coding for gaussian channels," *Information Theory, IEEE Transactions on*, vol. 58, no. 2, pp. 530–547, 2012.

[23] J. Perry, P. A. Iannucci, K. E. Fleming, H. Balakrishnan, and D. Shah, "Spinal codes," in *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, ser. SIGCOMM '12. New York, NY, USA: ACM, 2012, pp. 49–60. [Online]. Available: http://doi.acm.org/10.1145/2342356.2342363

[24] H. Balakrishnan, P. Iannucci, J. Perry, and D. Shah, "De-randomizing shannon: The design and analysis of a capacity-achieving rateless code," *CoRR*, vol. abs/1206.0418, 2012.

[25] A. Gudipati and S. Katti, "Strider: automatic rate adaptation and collision handling," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 158–169, aug 2011. [Online]. Available: http://doi.acm.org/10.1145/2043164.2018455

[26] M. Luby, "Lt codes," in *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, 2002, pp. 271–280.

[27] A. Shokrollahi, "Raptor codes," *Information Theory, IEEE Transactions on*, vol. 52, no. 6, pp. 2551–2567, 2006.

[28] R. Palanki and J. S. Yedidia, "Rateless codes on noisy channels," in *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, 2004, pp. 37–.

[29] G. T. . 3rd Generation Partnership Project, "Technical specification group radio access network; evolved terrestrial radio access (e-utra); user equipment (ue) radio transmission and reception," March 2009.

[30] P. Iannucci, J. Perry, H. Balakrishnan, and D. Shah, "No symbol left behind: A link-layer protocol for rateless codes," in *ACM MobiCom*, Istanbul, Turkey, August 2012.

[31] T. Villa, R. Knopp, and R. Merz, "Dynamic resource allocation for time-varying channels in next generation cellular networks part II: Applications in LTE," 2013.

[32] G. T. . 3rd Generation Partnership Project, "Technical specification group radio access network; evolved terrestrial radio access (e-utra); physical layer procedures," March 2011.

[33] E. Dahlman, S. Parkvall, and J. Sköld, *4G LTE/LTE-Advanced for Mobile Broadband*, 1st ed. Academic Press, 2011.

[34] A. Larmo, M. Lindström, M. Meyer, G. Pelletier, J. Torsner, and H. Wiemann, "The LTE link-layer design," *IEEE Communications Magazine*, pp. 52–59, April 2009.

[35] S. G. Wilson, *Digital Modulation and Coding*. Prentice Hall, 1996.

[36] T. Tabet and R. Knopp, "Cross-layer based analysis of multi-hop wireless networks," *IEEE Transactions on Communications*, vol. 58, no. 7, July 2010.

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

[38] A. Annamalai and C. Tellambura, "A simple exponential integral representation of the generalized marcum q-function qm (a, b) for real-order m with applications," in *Military Communications Conference, 2008. MILCOM 2008. IEEE*, nov. 2008, pp. 1 –7.

[39] N. G. Shephard, "From characteristic function to distribution function: A simple framework for the theory," *Econometric Theory*, no. 7, pp. 519–529, 1991.

[40] T. Villa, "Dynamic resource allocation for cellular networks with interference," Ph.D. dissertation, Telecom ParisTech, 2013.