# Adaptive HARQ with Non-Binary Repetition Coding

Stephan Pfletschinger, *Senior Member IEEE*, David Declercq, *Senior Member IEEE*, and Monica Navarro, *Senior Member IEEE*

***Abstract*—We consider Incremental Redundancy Hybrid Automatic Repeat reQuest (IR-HARQ) in which the code rate and modulation of the initial transmission and all retransmissions are adjusted based on average channel statistics. In the absence of instantaneous channel state information at the transmitter (CSIT), we present a method which computes, prior to transmission, the optimum code rates and modulations and explicitly consider a given maximum number of retransmissions. For the case that additional feedback on CSI of previous transmission attempts is available, we present two heuristic schemes which exploit this knowledge and offer increased throughput at the cost of higher computational complexity. We employ a rate-adaptive non-binary LDPC coding scheme which makes use of non-binary repetitions. While this coding scheme is particularly well-suited for adaptive IR-HARQ, we note that the presented analysis can be applied to any other channel code which employs soft decoding.**

***Index Terms*—HARQ, incremental redundancy, non-binary LDPC codes**

## I. Introduction

THE principle of adaptive transmissions is essential to wireless and mobile communication systems from their inception. Due to the time-varying nature of the radio channel and of the data traffic, adaptation is present in modern wireless systems at various levels: scheduling and resource allocation at the MAC layer, and power control and link adaption at the physical layer. The latter is a single-user procedure which aims to optimize the transmitted signal in order to obtain the best possible QoS parameters, typically given by data rate and error probability. This procedure is realized as the selection of a proper Modulation and Coding Scheme (MCS) out of a predefined set. Accurate adaptation is difficult if no or only outdated CSI is available at the transmitter, leading to a significant rate reduction if the error probability cannot exceed a given threshold. In this situation, transmission mechanisms with fast link-layer retransmissions come into play: modern HARQ protocols, implemented with Chase Combining (CC), Incremental Redundancy (IR) or a combination of both, can cope well with imperfect CSI at the Transmitter (CSIT) and achieve a throughput close to the perfect CSIT case.

Adaptive Coding and Modulation (ACM) algorithms are often specified without accounting for the effect of HARQ which is seen as a safety net rather than an integral part of the link layer. However, what is relevant for higher protocol layers are the throughput and the packet error probability *after* the link-level HARQ, relegating the bit or packet error rate at the initial transmission to be a secondary parameter. This is recognized e.g. in WiMAX [1] which recommends to optimize the link adaptation algorithm for the performance *after* the HARQ process. Further improvements are possible by adapting the code and modulation parameters for each retransmission – a possibility which is foreseen in LTE [2]–[4] and whose specific implementation is left to the manufacturers.

From the theoretical side, a throughput analysis of HARQ protocols for the slotted multiple-access channel under idealized but fairly general conditions was presented by Caire and Tuninetti [5], which assured the usefulness of combined coding and retransmissions for bursty packet transmissions. This work laid the ground for the analytical throughput evaluation of IR-HARQ with binary LDPC codes in the asymptotic regime by Sesia et al. [6]. In [7], Negi and Cioffi developed an optimum power allocation strategy for the delay-constrained capacity of the block fading channel with causal feedback. Considering type-I HARQ and constraints on the packet buffer, Djonin et al. presented a control-theoretic framework for rate and power adaptation [8]. On the other hand, Wu and Jindal confirmed the benefits of HARQ in block fading channels and showed that the throughput comes close to the ergodic capacity even with few retransmissions [9]. While Wu and Jindal base their analysis on outage probability and information-theoretic tools, Lagrange [10] derives an approximate formula for the throughput of CC-HARQ based on an analytical approximation of the packet error probability for a given MCS. An extension to cooperative HARQ and correlated fading is given by Harsini et al. [11], which employ a Markov model for the temporal correlation of successive transmissions but apply the same MCS in all retransmissions.

While these theoretical analyses focus on fixed retransmission parameters, assuming the same modulation and block length for all transmissions, allowing for adaptive retransmission parameters does not incur any major complication to a system which already employs ACM. Several recent works consider adaptive retransmissions, e.g. [12]–[17], and obtain the common conclusion that adapting the retransmission parameters is beneficial in terms of throughput and average delay. A difficulty lies in the optimization of the transmission parameters for truncated HARQ, i.e. for a limited – and typically small – number of retransmissions, with practical modulation and coding schemes. A fairly general framework for the analysis of various HARQ protocols has been introduced by Cheng [18], which proposes the *ACcumulated*

S. Pfletschinger and M. Navarro are with the Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Av. Carl Friedrich Gauss 7, 08860 Castelldefels, Spain (e-mails: {stephan.pfletschinger, monica.navarro}@cttc.es).

D. Declercq is with the ETIS laboratory, ENSEA/univ. Cergy-Pontoise/CNRS-UMR-8051, 6, Avenue du Ponceau, F-95000, Cergy-Pontoise, France (e-mail: declercq@ensea.fr).

*Mutual Information (ACMI)* instead of the average SNR as a unified metric. In [13], the retransmission parameters are adapted by a genetic algorithm for a turbo-coded system while in [14], [15] the performance is described in terms of outage probability and a numerical optimization of the retransmission rates is carried out to maximize the throughput. Both approaches reflect the difficulties in accurately describing a realistic setting and obtaining a feasible adaptation algorithm. Despite the assumption of an idealized coding scheme in [14], [15], the results provide valuable guidance for practical schemes employing good channel codes. Another approach for adapting the size of the retransmission units has been presented by Visotsky et al. [19]. That work presents an adaptive scheme which relies on the CSI of previous transmissions and is valid for convolutional coding and binary modulation.

In this work, we analyze the performance of variable-rate truncated HARQ and propose allocation and adaptation methods which maximize the throughput subject to a target error rate for the block fading channel:

- Rate *allocation* precomputes the modulations and code rates for the initial transmission and for $L-1$ possible retransmissions in an *offline* manner, based solely on the statistics of the block fading channel. This scheme only requires a one-bit feedback in form of an ACK/NACK message after each transmission.
- Rate *adaptation* exploits feedback on previously failed transmission attempts and finds in an *online* manner the optimum code rate and modulation for the next retransmission. For this adaptation, feedback on the received mutual information of the previous transmission is required. We present two heuristic approaches which both achieve higher throughput than the rate allocation scheme.

One of the main differences with previous work is that we consider a practical coding scheme with finite block lengths, for which the asymptotic approximation with outage probabilities is inaccurate and we explicitely consider adjusting the modulation for each retransmission. As a state-of-the-art coding scheme, we employ a Non-Binary (NB) LDPC code which features a non-binary repetition scheme that achieves a coding gain without increasing the decoding complexity [20]. While this channel code is particularly well-suited to adaptive HARQ, we note that other coding schemes which apply soft decoding can be treated with the same methodology.

The rest of the paper is organized as follows: Section II introduces the system model and the rate-adaptive coding and modulation scheme, while Section III analyzes its performance based on the notion of accumulated mutual information. Secion IV describes the proposed adaptation strategies for throughput maximization. In the Appendix, a possible extension towards partial CSIT is given.

## II. System Model and Non-Binary Coding Scheme

### A. Non-Binary Repetition Coding, Puncturing and Adaptive Modulation

*1) Non-Binary LDPC Codes and Decoder:* An LDPC code is defined as the set of all codewords $\mathbf{c} = [c_1, c_2, \ldots, c_{\bar{N}}]$ which satisfy the parity-check equation

$$\mathbf{H}\mathbf{c}^{\mathrm{T}} = \mathbf{0}, \tag{1}$$

where $\mathbf{H}$ is a *sparse* $\bar{M} \times \bar{N}$ matrix. For non-binary LDPC codes, the parity-check equation is defined over the Galois field $\mathbb{F}_q = \{\alpha_0, \alpha_1, \ldots, \alpha_{q-1}\}$, where $\alpha_0 = 0$ and $\alpha_1 = 1$ denote the additive and multiplicative identity, respectively. The field order $q$ is typically a power of two and $q = 2$ includes the particular binary case. For full-rank matrices, $K = \bar{N} - M$ is the length of the information block to be encoded and the code rate is given by $R_{\mathrm{m}} = K/\bar{N}$. In this paper, we choose the particular family of regular NB-LDPC codes with a constant column weight of two. These codes are called *ultra-sparse* or *cycle* NB-LDPC codes [21] and are known to have the best performance at high SNR for a field order $q \geq 64$ [22]. Moreover, these codes can be efficiently designed for finite lengths and efficient encoding algorithms are available [23], [24].

NB-LDPC codes are often proposed as an alternative to their binary counterparts for small to moderate block lengths or when the channel cannot be represented as a binary-input memoryless channel, which is the case for higher-order modulations and for multiple transmit antennas [25], [26]. The performance improvement of NB-LDPC codes comes at the price of an increased decoding complexity, but recently many works have demonstrated that the performance gains of NB-LDPC codes remain valid, even with low-complexity, sub-optimal decoders [27]–[30].

For coded modulation, before transmission, the code symbols $c_n \in \mathbb{F}_q$ are mapped to QAM symbols, while at the receiver side the received channel symbols $\mathbf{y}_n \in \mathbb{C}^T$ are demapped to a Log-Likelihood Ratio (LLR) vector per code symbol. The received symbols $\mathbf{y}_n$ might be scalar or vectorial, i.e. $T = 1, 2, \ldots$, depending on the modulation and the channel model, as detailed in Subsection II-A3. The LLR vector corresponding to $c_n$ is given by $\mathbf{C}_n = [C_{n,0}, C_{n,1}, \ldots, C_{n,q-1}]^{\mathrm{T}}$ with

$$C_{n,g} \triangleq \ln \frac{P\left[c_n = \alpha_g \mid \mathbf{y}_n\right]}{P\left[c_n = \alpha_0 \mid \mathbf{y}_n\right]} \text{ for } \begin{array}{l} n = 1, 2, \ldots, \bar{N} \\ g = 0, 1, \ldots, q-1 \end{array}, \tag{2}$$

where $g \in \{0, 1, \ldots, q-1\} \subset \mathbb{N}$ denotes an integer number which serves as index to the GF element $\alpha_g \in \mathbb{F}_q$. We note that, although $C_{n,g}$ according to (2) is actually the logarithm of the ratio of *a posteriori* probabilities, we use the term *LLR* as it is well established in the literature of soft and iterative decoding. In the same direction, the division by $P[c_n = \alpha_0 \mid \mathbf{y}_n]$ in (2) is introduced for consistency with literature (see e.g. [31] for an introduction to *Log-Likelihood algebra*).

Let us now describe our protograph-based code design. First introduced by [32], a binary protograph is defined as a small bipartite graph from which a larger graph is obtained by the so-called *lifting* technique. The protograph itself is generally described by its *adjacency* or *base* matrix $H_{\mathrm{B}}$, where the coefficients $H_{\mathrm{B}}(m, n)$ represent the number of edges between the $m$-th check node of the protograph and the $n$-th variable node [33]. The base matrix $H_{\mathrm{B}}$ is hence a small matrix containing small integer values. The lifting operation expands the base matrix by replacing each nonzero entry $H_{\mathrm{B}}(m, n) > 0$ by the

same number of non-overlapping circulant matrices. Circulant matrices are usually preferred for practical purposes since this reduces the descriptional complexity (ie. storage) of the parity check matrix in the hardware realizations of the LDPC encoder and decoder. If $L_{\mathrm{c}}$ is the size of the circulant matrices, we obtain after lifting a Tanner graph with $L_{\mathrm{c}}$ times more nodes and edges than the protograph. The last step for non-binary LDPC codes is then to assign nonzero values to the edges of the lifted Tanner graph. For this step, we follow the procedure described in [21].

In Fig. 1, we show the protograph which has been chosen to define the mother code with code rate $R_{\mathrm{m}} = 1/2$. The structure of the protograph has been chosen so as to maximize the number of *one-step-recovery* (1-SR) survivors [34], [35] in order to improve the performance of punctured schemes, as described in the next section.

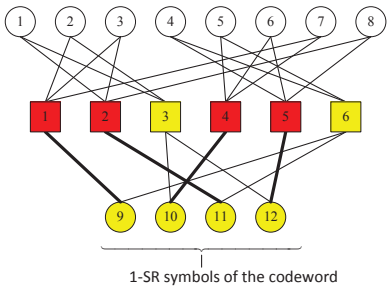


Figure 1. Detailed protograph optimized for puncturing. This protograph has the property to maximize the number of symbols with the 1-SR property.

*2) Puncturing and Non-Binary Repetition Coding:* For adaptive HARQ, we require a wide range of code rates which are obtained by either puncturing a mother code or extending the same mother code with additional symbols obtained from non-binary multiplication coding. Puncturing is a well-known technique for increasing the code rate of a given mother code while non-binary multiplication coding is a novel scheme, well adapted to NB-LDPC codes, and which allows to derive lower rates from a given code.

In [34], the authors proposed a criterion for deriving puncturing patterns from the knowledge of the Tanner graph properties. They introduced the concept of a $k$-SR survivor symbol, which is defined as a symbol which can be recovered from the other symbols in its Tanner graph neighborhood after $k$ iterations of the message passing decoder, assuming that the other symbols are correctly decoded. For this reason, 1-SR survivor symbols are preferably punctured and with the protograph in Fig. 1 we have four of these symbols, which is the maximum possible number for code rate $R_{\mathrm{m}} = 1/2$. For more details on the design of the protograph, we refer to [36].

Recently, another important benefit of NB-LDPC codes has been identified in [37]: the non-binary multiplication coding can be seen as a flexible non-binary repetition scheme with simple Galois field multiplications that can achieve a significant coding gain over the usual binary repetition coding while the decoding complexity is hardly increased. This gain is particularly pronounced for higher field orders with $q \geq 64$.

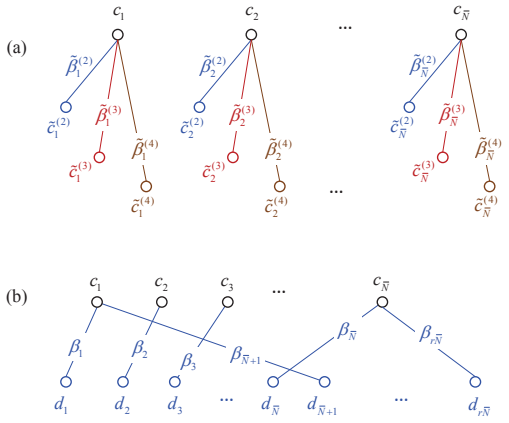


Figure 2. Two representations of multiplicative repetition coding. In both cases, the symbols $\tilde{c}_n^{(j)}$ and $d_n$, respectively, are obtained by multiplication of one mother codeword symbol $c_n$.

Starting from a mother codeword $\mathbf{c} = [c_1, c_2, \ldots, c_{\bar{N}}] \in \mathbb{F}_q^{\bar{N}}$, additional symbols $\tilde{c}_n^{(j)}$ can be generated in a *fountain-like* fashion by a simple multiplication in $\mathbb{F}_q$ as

$$\tilde{c}_n^{(j)} = \tilde{\beta}_n^{(j)} \cdot c_n \text{ for } j = 1, \ldots, r, \tag{3}$$

where $r$ denotes the maximum number of repeated symbols and we set $\tilde{\beta}_n^{(1)} = 1$ for the mother codeword. This process is depicted in Fig. 2(a), where each edge in the graph corresponds to a multiplier $\tilde{\beta}_n^{(j)}$. For $\tilde{\beta}_n^{(j)} = 1$ for $j > 1$, we obtain the standard repetition coding, in which the codeword symbols $c_n$ are repeated unaltered. In order to keep the encoder simple, we focus on repetition schemes with a multiplier $\tilde{\beta}^{(j)}$ that does not depend on the codeword symbol index $n$ but only on the repetition index $j = 1, \ldots, r$. By proper optimization, this limitation does not cause a performance loss [36].

Decoding for this non-binary repetition scheme is particularly simple and hardly more complex than for binary repetitions: According to the definition of the LLR-vector (2) and assuming an memoryless channel, i.e. $p(\mathbf{y} \mid \mathbf{c}) = \prod_n p\left(\mathbf{y}_n^{(j)} \mid \tilde{c}_n^{(j)}\right)$, in addition to equally probable values of the codeword symbols, i.e. $P[c_n = \alpha_g] = 1/q$, we obtain

$$\begin{aligned} C_{n,g} &= \ln \frac{\prod_{j=1}^{r} p\left(\mathbf{y}_n^{(j)} \mid \tilde{c}_n^{(j)} = \tilde{\beta}_n^{(j)} \alpha_g\right)}{\prod_{j=1}^{r} p\left(\mathbf{y}_n^{(j)} \mid \tilde{c}_n^{(j)} = \tilde{\beta}_n^{(j)} \alpha_0\right)} \\ &= \sum_{j=1}^{r} \tilde{C}^{(j)}_{n,[\tilde{\beta}_n^{(j)} \alpha_g]}, \end{aligned} \tag{4}$$

where the product $\tilde{\beta}_n^{(j)} \alpha_g$ is performed in the Galois field $\mathbb{F}_q$ and $[\alpha_g] = g$ denotes the index of the GF element $\alpha_g$. The LLR values of the repeated symbols are defined in the same way as for the original code symbols, i.e.

$$\tilde{C}_{n,g}^{(j)} \triangleq \ln \frac{P\left[\tilde{c}_n^{(j)} = \alpha_g \mid \mathbf{y}_n^{(j)}\right]}{P\left[\tilde{c}_n^{(j)} = \alpha_0 \mid \mathbf{y}_n^{(j)}\right]} \tag{5}$$

and $\mathbf{y}_n^{(j)}$denotes the received channel symbol for $\tilde{c}_n^{(j)}$. Note that $\tilde{\beta}_n^{(j)} \cdot \alpha_0 = \alpha_0 = 0$. Decoding hence amounts to a simple summation of the LLR vectors that are associated with each code symbol. This operation is transparent to the decoder of the mother NB-LDPC code, which therefore does not require any modification.

While the generation of repetition symbols according to this procedure conveys the basic idea of non-binary repetition coding and the corresponding decoding, for the purpose of HARQ, the description of multiplicative repetition and puncturing depicted in Fig. 2(b) is more convenient. We define a long mother codeword $\mathbf{d} \triangleq [d_1, d_2, \ldots, d_{r\bar{N}}] \in \mathbb{F}_q^{r\bar{N}}$ with

$$d_n = \beta_n \cdot c_{\pi(n)}, \quad \text{for } n = 1, 2, \ldots, r\bar{N} \tag{6}$$

where $\pi : \{1, 2, \ldots, r\bar{N}\} \to \{1, 2, \ldots, \bar{N}\}$ is an index mapping defined by the repetition scheme above. This description is an equivalent alternative to (3) which also specifies the puncturing by defining the transmission order of the first $\bar{N}$ symbols. The first $\bar{N}$ symbols in $\mathbf{d}$ are the same as the ones in $\mathbf{c}$, but their order is given by the puncturing scheme.

At the receiver, first the LLR-vectors for the symbols $d_n$ are computed according to

$$D_{n,g} = \ln \frac{P\left[d_n = \alpha_g \mid \mathbf{y}_n\right]}{P\left[d_n = \alpha_0 \mid \mathbf{y}_n\right]}, \; n = 1, 2, \ldots, r\bar{N} \tag{7}$$

and then combined to

$$C_{n,g} = \sum_{m:\, \pi(m)=n} D_{m,[\beta_m \cdot \alpha_g]} \qquad \begin{matrix} n = 1, 2, \ldots, \bar{N} \\ g = 0, 1, \ldots, q-1 \end{matrix}, \tag{8}$$

where $\beta_i \cdot \alpha_g$ is again the multiplication in $\mathbb{F}_q$ and $[\alpha_g] = g$ denotes the index of $\alpha_g \in \mathbb{F}_q$.

A simple method for implementing this summation is described in Algorithm 1. This LLR-value combination is of low complexity and becomes identical to maximum ratio combining for binary repetition coding with BPSK or QPSK modulation.

**Algorithm 1** Combination of LLR-values

```
C_{n,g} = 0 ∀n, g
for n = 1, 2, ..., rN̄ do
    for g = 0, 1, ..., q − 1 do
        C_{π(n),g} = C_{π(n),g} + D_{n,[β_n · α_g]}
    end for
end for
```

*3) Adaptive Modulation for q-ary Channel Coding:* While with $q$-ary channel codes, we can in principle apply the usual $M$-QAM constellations [38] in a similar way as for binary codes, we can take advantage of the higher field order: For $q = M^T$, we can map *one* codeword symbol $c_n \in \mathbb{F}_q$ to $T \in \mathbb{N}$ channel uses. Let us define the mapping

$$\boldsymbol{\mu}_T : \mathbb{F}_q \to \mathcal{A}_T \subset \mathbb{C}^T, \tag{9}$$

where $\mathcal{A}_T = \{\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{q-1}\} \subset \mathbb{C}^{\mathrm{T}}$ denotes a constellation of $q$ points in $T$ complex dimensions and we define the mapping straightforwardly as

$$\boldsymbol{\mu}_T(\alpha_g) = \mathbf{a}_g \quad \text{for } g = 0, 1, \ldots, q-1. \tag{10}$$

This QAM symbol is transmitted over a flat fading channel and received as

$$\mathbf{y}_n = h \cdot \boldsymbol{\mu}_T(d_n) + \mathbf{w}_n, \; \mathbf{w}_n \sim \mathcal{CN}\left(0, N_0 \mathbf{I}_T\right). \tag{11}$$

In the simplest case, multidimensional constellations $\mathcal{A}_T$ for $T > 1$ are obtained as the Cartesian product of the usual QAM constellations. In the following, we will focus on a field size of $q = 256$ in combination with 256-QAM and 16-QAM, which correspond to $T = 1$ and $T = 2$, respectively. For other values of $T$, multidimensional constellations can be derived from $2T$-dimensional sphere packings [39], as described e.g. in [26].

The LLR-values for the flat fading channel (11) with perfect CSI at the receiver follows directly from the definition (7) by applying Bayes' theorem and is given by

$$D_{n,g} = -\frac{|\mathbf{y}_n - h\mathbf{a}_g|^2}{N_0} + \frac{|\mathbf{y}_n - h\mathbf{a}_0|^2}{N_0}. \tag{12}$$

This expression involves no marginalization and is therefore of much lower complexity than in the case of binary LLR-values with higher-order modulation. For $T = 1$, the involved signals reduce to scalars.

This approach has two advantages compared to mappings which involve several codeword symbols:

1) The Bayesian optimum soft demapper has low complexity.
2) If the physical channel is memoryless, the equivalent channel seen by the coding scheme remains memoryless. This is a condition which is assumed by the BP decoder.

For binary coding with higher-order modulation, but also for non-binary coding with $M^T \neq q$ (see e.g. the case of $M = 16$, $q = 64$ in [38]), these conditions are not satisfied.

### *B. HARQ Model and Channel Model*

The information to be transmitted is represented by the message $\mathbf{u} \in \mathbb{F}_q^K$ containing $K$ symbols in the Galois field $\mathbb{F}_q$ of order $q$, corresponding to $K_{\mathrm{bin}} = \mathrm{ld}q \cdot K$ bits, where $\mathrm{ld}q = \log_2 q$ denotes the base-2 logarithm ("logarithmus dualis"). This message is encoded to produce a codeword of the mother NB-LDPC code, $\mathbf{c} = [c_1, c_2, \ldots, c_{\bar{N}}] \in \mathbb{F}_q^{\bar{N}}$. From this codeword, up to $L$ codeblocks $\mathbf{d}_\ell$ are formed by multiplicative repetition and possibly by puncturing. These blocks are modulated by possibly different modulations and are transmitted in subsequent HARQ rounds, denoting $\mathbf{d}_1$ the block of codeword symbols of the initial transmission.

For the description of the HARQ process, it is convenient to assume that a long codeword $\mathbf{d} = [d_1, \ldots, d_{r\bar{N}}]$ according to (6) has been computed by the encoder. From this long mother codeword $\mathbf{d}$ we can derive higher-rate code blocks $\mathbf{d}_\ell = \left[d_{\tilde{N}_{\ell-1}+1}, d_{\tilde{N}_{\ell-1}+2}, \ldots, d_{\tilde{N}_\ell}\right]$ of lengths $N_\ell$, where we set $\tilde{N}_\ell \triangleq \sum_{i=0}^{\ell} N_i$ with $N_0 = 0$ and it must hold $\tilde{N}_L \leq r\bar{N}$. This corresponds to rate-compatible puncturing [40] of the long codeword $\mathbf{d}$ and, by setting $\beta_n = 1 \; \forall n$, it includes binary repetition coding as a special case. As illustrated in Fig. 3, the code blocks $\mathbf{d}_\ell$ are modulated to the symbol sequences $\mathbf{x}_\ell \in \mathbb{C}^{N_\ell T_\ell}$ which are given by

$$\mathbf{x}_\ell = \left[\boldsymbol{\mu}_{T_\ell}\left(d_{\tilde{N}_{\ell-1}+1}\right), \ldots, \boldsymbol{\mu}_{T_\ell}\left(d_{\tilde{N}_\ell}\right)\right] \tag{13}$$

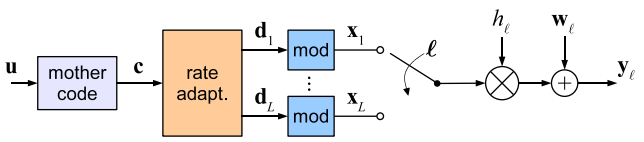


Figure 3. Transmitter and channel model for truncated HARQ with $L$ transmissions

and $T_\ell$ denotes the number of channel uses per codeword symbol at the $\ell$-th transmission. These sequences are transmitted in the $\ell$-th transmission and are received as

$$\mathbf{y}_\ell = h_\ell \cdot \mathbf{x}_\ell + \mathbf{w}_\ell, \quad \ell = 1, 2, \ldots, L, \tag{14}$$

where the fading coefficient $h_\ell$ is i.i.d. Rayleigh distributed and constant during one code block, i.e. $h_\ell \sim \mathcal{CN}(0,1)$, while the noise is white Gaussian, $\mathbf{w} \sim \mathcal{CN}(0, N_0 \mathbf{I}_{N_\ell})$. This model is similar to the one applied in [18], with the difference that we additionally allow to adapt the modulation per code block.

We denote by $\mathtt{E}_\ell$ the event that decoding fails in the $\ell$-th transmission, while $\bar{\mathtt{E}}_\ell$ denotes correct decoding in the $\ell$-th transmission. We assume that all decoding errors are detected, which is a light idealization of the inherent error detection capability of an LDPC code or an additional CRC code for error detection. We define as in [13], [14]

$$s_\ell \triangleq P\left[\mathtt{E}_1, \mathtt{E}_2, \ldots, \mathtt{E}_{\ell-1}, \bar{\mathtt{E}}_\ell\right] \text{ success in slot } \ell \tag{15}$$

$$f_\ell \triangleq P\left[\mathtt{E}_1, \mathtt{E}_2, \ldots, \mathtt{E}_{\ell-1}, \mathtt{E}_\ell\right] \ \ell \text{ failures.} \tag{16}$$

It holds $s_\ell = f_{\ell-1} - f_\ell$ and we set $f_0 = 1$. The throughput and the average number of transmissions are given by [6]

$$\eta = \frac{K_{\text{bin}}\,(1 - f_L)}{\sum_{\ell=1}^{L} f_{\ell-1} N_\ell T_\ell} \tag{17}$$

$$\tau = \frac{\sum_{\ell=0}^{L-1} f_\ell}{1 - f_L} \tag{18}$$

Here, the throughput is measured as the average number of bits per channel use, which corresponds to the spectral efficiency in $\frac{\text{bit/s}}{\text{Hz}}$.

The variable lengths of the codewords leads to implications at the system level which have to be considered in the desgin of the complete communication system. For multiple-access systems, e.g. OFDMA, an informed scheduler might assign exactly the required resources and avoid, as far as possible, unused time-frequency slots. On the other hand, it is also possible to combine the transmission of multiple (partial) codewords in the same time slot and thus avoid unused resources. The most appropriate strategy to deal with these variable-length codewords depends on the link layer and the multiple-access scheme and is outside the scope of this paper. See also the brief discussion on this topic in [15].

## III. Numerical Analysis

For all numerical results in this paper, we use a non-binary LDPC code of rate $R_{\text{m}} = 1/2$, field order $q = 256$ and message length $K = 90$. The message or packet length in bits is thus $K_{\text{bin}} = \text{ld}q \cdot K = 720$ bits and the codeword length is $N_{\text{m}} = 180$ symbols. From this mother code, higher and lower rates are derived by puncturing and non-binary repetition with $r_{\max} = 7$. In the following, we will apply the word lengths

$$N \in \mathcal{N}_{\text{c}} = \{100, 108, 120, 135, 150, 180, 220, 270, 360, 450, \ldots, 1080, 1170, 1260\}. \tag{19}$$

This means that with puncturing, the commonly used code rates $\frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{5}{6}, \frac{9}{10}$ are obtained, while multiplicative repetition yields the code rates $\frac{9}{22}, \frac{1}{3}, \frac{1}{4}, \ldots, \frac{1}{14}$ .

### A. Performance of the Coding Scheme as a Function of the Accumulated Mutual Information

We apply the modulations 256-QAM and 16-QAM, which correspond to $T = 1$ and $T = 2$ channel uses per codeword symbol. For 256-QAM over an AWGN channel, we obtain the set of Packet Error Rate (PER) curves plotted in Fig. 4, where we can observe that with the same modulation and adaptation of the code rate alone we can cover an SNR range of more than 25 dB.

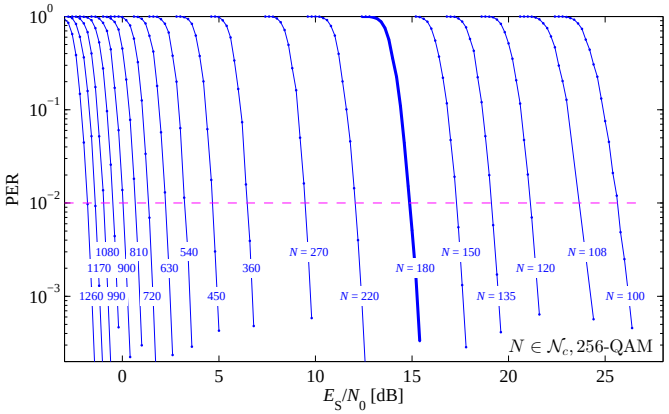


Figure 4. PER as a function of the SNR for all codeword lengths $N \in \mathcal{N}_{\text{c}}$ and for 256-QAM ($T = 1$) over the AWGN channel. The bold curve corresponds to the mother code.

In order to show the gain of the non-binary repetition scheme compared to binary repetition, we can plot the SNR-rate points for a fixed target PER, as shown in Fig. 5, where the rate is defined as $R = \frac{K_{\text{bin}}}{TN}$. In addition, the values for binary repetition of the mother codeword are plotted: let $\gamma_{\text{m}}$ denote the required SNR to achieve the target PER for the mother code of rate $R_{\text{m}} = 1/2$; then the required SNR for an $r$-fold repetition is given by $\gamma_{\text{m}}/r$. We can observe that the required SNR using the multiplicative NB-repetition coding is significantly superior and comes close to the capacity curve.

In order to characterize the PER in an HARQ scheme with incremental redundancy, we need to relate the SNR of the channel to the *accumulated mutual information* (ACMI) experienced by the coding scheme [18].

The mutual information per codeword symbol is given by

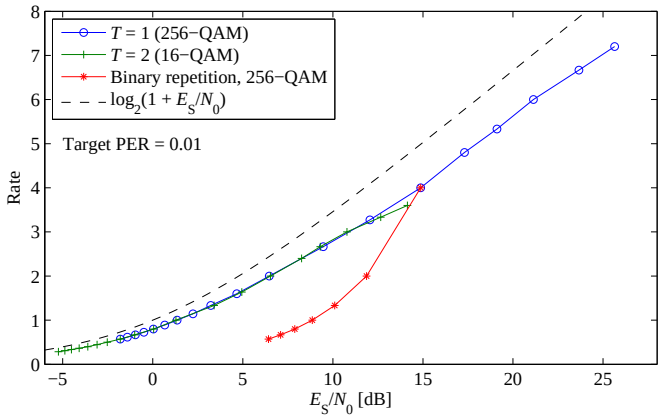


Figure 5. Rate-SNR pairs at $\mathrm{PER} = 0.01$ for all $N \in \mathcal{N}_\mathrm{c}$

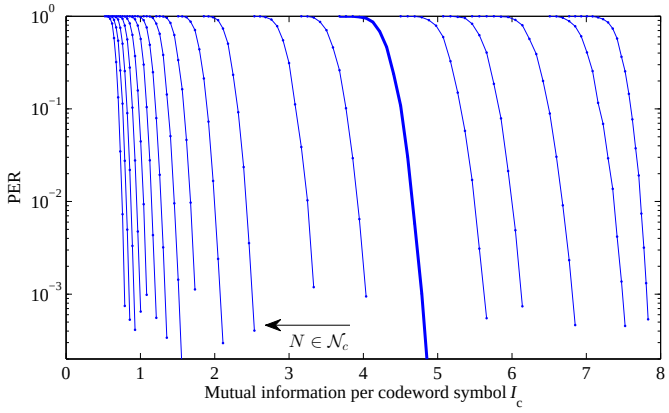


Figure 6. PER as a function of the mutual information per codeword symbol for $N \in \mathcal{N}_\mathrm{c}$. The bold curve corresponds to the mother code.

the Coded Modulation (CM) capacity [41]

$$I_\mathrm{c}(\gamma) = I\left(c;\mathbf{y}\right) = \mathrm{ld}q - \frac{1}{q}\sum_{i=0}^{q-1}\mathbb{E}_\mathbf{y}\left[\mathrm{ld}\frac{\sum_{j=0}^{q-1} p\left(\mathbf{y}\mid\mathbf{a}_j\right)}{p\left(\mathbf{y}\mid\mathbf{a}_i\right)}\right]. \quad (20)$$

Although there exists no closed-form expression, this mutual information can be computed easily by Monte-Carlo integration.

According to the block fading model, the channel in the $\ell$-th transmission appears as an AWGN channel with a constant SNR $\gamma_\ell$ and the corresponding mutual information $I_\ell = I_\mathrm{c}\left(\gamma_\ell\right)$. The ACMI after $\ell$ transmissions is hence given by $I_\mathrm{w} = \sum_{i=1}^{\ell} I_i N_i$.

It has been observed by many authors that the PER of an LDPC code is approximately determined by the mutual information of the channel between the encoder and decoder [42]. This observation has been applied for the definition of link-to-system interfaces in simulators [43] and can also be applied here to define the function $p_\mathrm{w}\left(I_\mathrm{w}, N\right)$, which relates the ACMI $I_\mathrm{w}$ and the codeword length $N$ to the PER, independently of the modulation index $T$. This function can be evaluated numerically on the basis of a look-up-table of the simulation results and is plotted in Fig. 6. Note that since the packet (message) length is fixed to $K = 90$ symbols, the codeword length $N$ determines the code rate as $K/N$, whereas the mutual information per codeword symbol is $I_\mathrm{c} = I_\mathrm{w}/N$.

### B. Error Probabilities in HARQ

According to the channel model defined by (14), the *instantaneous* SNR at the $\ell$-th transmission is given by

$$\gamma_\ell = \frac{|h_\ell|^2 E_\mathrm{S}}{N_0} = |h_\ell|^2\bar{\gamma}, \quad (21)$$

where $\bar{\gamma} = E_\mathrm{S}/N_0$ denotes the *average* SNR. For the Rayleigh fading channel, the instantaneous SNR is exponentially distributed with probability density function (pdf) $f_\gamma(x) = \frac{1}{\bar{\gamma}}\exp\left(-\frac{x}{\bar{\gamma}}\right)$ for $x \geq 0$. We denote the pdf of the mutual information per codeword symbol by $f_I(x)$, with $x \in [0, \mathrm{ld}q]$. This function cannot be expressed in closed-form as there is no closed-form relation between the SNR and the mutual information (20). On the other hand, if (20) is given numerically, it is straightforward to generate samples of a random variable with the desired pdf $f_I(x)$.

With the function $p_\mathrm{w}\left(\sum_{i=1}^{\ell} I_i N_i, \sum_{i=1}^{\ell} N_i\right)$ denoting the probability of error at the $\ell$-th transmission, the probability of $\ell$ consecutive failures, defined by (16), can be expressed as

$$\begin{aligned} f_\ell &= \int_0^{\mathrm{ld}q}\cdots\int_0^{\mathrm{ld}q}\int_0^{\mathrm{ld}q} p_\mathrm{w}\left(I_1N_1, N_1\right) \\ &\quad\cdot p_\mathrm{w}\left(I_1N_1 + I_2N_2, N_1 + N_2\right)\cdots p_\mathrm{w}\left(\sum_{i=1}^{\ell} I_iN_i, \sum_{i=1}^{\ell} N_i\right) \\ &\qquad\cdot f_I(I_1)f_I(I_2)\cdots f_I(I_\ell)\,\mathrm{d}I_1\mathrm{d}I_2\cdots\mathrm{d}I_\ell \\ &= \mathbb{E}_{I_1\cdots I_\ell}\Big[p_\mathrm{w}\left(I_1N_1, N_1\right)\cdot p_\mathrm{w}\left(I_1N_1 + I_2N_2, N_1 + N_2\right) \\ &\qquad\cdots p_\mathrm{w}\left(\sum_{i=1}^{\ell} I_iN_i, \sum_{i=1}^{\ell} N_i\right)\Big] \\ &= \mathbb{E}_{I_1\cdots I_\ell}\left[\prod_{j=1}^{\ell} p_\mathrm{w}\left(\sum_{i=1}^{j} I_iN_i, \sum_{i=1}^{j} N_i\right)\right]. \end{aligned} \quad (22)$$

From the last expression, we see that the probabilities $f_\ell$ can be evaluated numerically by Monte-Carlo integration. For small values of $\ell$, this allows for a relatively simple numerical evalution. On the other hand, for large $\ell$, the Gaussian approximation as applied in [6] is more convenient. Note that the numerical calculation is not limited to a certain coding scheme or fading model, but can be applied in the same manner to any fading distribution and coding scheme which is characterized by a function $p_\mathrm{w}\left(I_\mathrm{w}, N\right)$.

For an information-theoretic evaluation, instead of a look-up table for a specific coding scheme, one could take a more general, semianalytical approach by applying a bound on the error rate in the finite block length regime according to Polyanskiy et al. [44]. While we do not follow this research line in this paper, we note that this approach might lead to

valuable results for finite block lengths.

## IV. Adaptive HARQ

In adaptive HARQ, we can adapt in each of the up to $L$ transmissions the length $N_\ell$ of the code block as well as the modulation index $T_\ell$. The initial code rate is given by $K/N_1$ and it is feasible to select different block lengths and modulations for the retransmissions. In a system which already employs adaptive coding and modulation, the introduction of adaptive retransmission units hardly increases the complexity.

### A. Optimization Criteria

An obvious optimization criterion is throughput maximization, while the delay is limited by the maximum number of retransmissions $L$. Due to the finite delay, it is not possible to achieve error-free communication and it is reasonable to target a residual word error probability $f_L > 0$. We can therefore formulate the optimization problem as

$$\max_{N_1,\cdots,N_L;T_1,\ldots,T_L} \eta \tag{23a}$$

$$\text{s.t. } f_L \leq \mathrm{PER}_{\max} \tag{23b}$$

where the throughput is defined by (17) and (16), (22), while the constraint on the number of transmissions is implicit in (17). We set the constraint on the PER in the following to $\mathrm{PER}_{\max} = 0.01$, noting that we obtain unconstrained throughput optimization by setting $\mathrm{PER}_{\max} = 1$. The problem (23) is a discrete optimization problem which can be solved numerically for small $L$ by an exhaustive search over all $N_\ell \in \mathcal{N}_\mathrm{c}$, $T_\ell \in \{1, 2\}$.

### B. Capacity Bounds

The achievable throughput is upper bounded by the ergodic capacity of the Rayleigh fading channel, which is given by

$$C_\mathrm{R} = \frac{1}{\ln 2} \exp\left(\frac{1}{\bar{\gamma}}\right) E_1\left(\frac{1}{\bar{\gamma}}\right), \tag{24}$$

where $E_1(x) \triangleq \int_x^\infty e^{-t}/t \, \mathrm{d}t$ is the exponential integral. While this expression assumes no constraints on the transmit signal other than a power constraint, we obtain a tighter bound by considering the modulation. In analogy to (20), the ergodic CM capacity is given by

$$C_\mathrm{CM} = \mathrm{ld} q - \frac{1}{q} \sum_{i=0}^{q-1} \mathbb{E}_{y,h} \left[ \mathrm{ld} \frac{\sum_{j=0}^{q-1} p\left(y \mid a_j, h\right)}{p\left(y \mid a_i, h\right)} \right], \tag{25}$$

where $q = 256$ and the symbols $a_i$ are taken out of a 256-QAM constellation.

### C. Reference Schemes: No Retransmission, Chase Combining and Incremental Redundancy

As a simple reference, we first consider a scheme without retransmissions, which is equivalent to the optimization problem (23) with $L = 1$. Without retransmissions and knowledge of only the average SNR $\bar{\gamma}$ at the transmitter, the constraint (23b) leads to a significant throughput reduction compared to the case where perfect CSIT is available [45], as can be observed in Fig. 7. For all throughput-SNR points in Fig. 7, at least $5 \cdot 10^4$ codewords have been simulated[1], hence the lack of smoothness of some curves is not due to an insufficient number of simulated codewords, but is rather caused by the limited number of possible code rates according to (19).

The best known reference scheme for HARQ is probably *Chase combining*: For the first transmission, one of the code rates $R_\mathrm{c} = \frac{K}{N_1} \in \left\{\frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{5}{6}, \frac{9}{10}\right\}$ is selected while for the retransmissions the same codeblock, with the same modulation, is repeated. The receiver applies maximum ratio combining of the received symbols. The obtained rates with this scheme for a maximum number of $L = 3$ transmissions are given in Fig. 7, where we can observe a significant performance gain compared to the scheme without retransmissions. It also interesting to note that the throughput obtained with the NB-LDPC code here comes close to the results derived in [46, Fig. 1] for a capacity-achieving code and for $L = 5$ transmissions.

The simplest scheme which exploits non-binary repetition coding is based on *Incremental Redundancy* (IR) with constant blocklengths and the same modulation for all transmissions. This scheme can therefore be compared directly to CC. Fixing again the maximum number of transmisssions to $L = 3$, the values for $N_1$ and $T_1$ are obtained by solving the simplified optimization problem (23) with the additional constraints

$$\begin{aligned} &N_1 = N_2 = N_3 \in \{100, 108, 120, 135, 150, 180\} \\ &T_1 = T_2 = T_3 \in \{1, 2\}. \end{aligned} \tag{26}$$

For a fair comparison with CC, we restricted the codeword length to values which correspond to the same set of code rate as for CC. The gain of IR with respect to CC is moderate as can be observed in Fig. 7, despite the significant advantage of non-binary repetition compared to binary repetition, which is equivalent to CC, in Fig. 5. However, this is not a contradiction since the gains for the AWGN channel are generally not reproduced in the block fading case and, more importantly, both schemes are identical in the initial transmission for code rates $R_\mathrm{c} \geq 1/2$.

### D. Rate Allocation: Offline Computation of Optimum Block Lengths and Modulations

Without the limitation of constant blocklengths, we can obtain the optimum values for the initial transmission and all retransmissions by solving the discrete optimization problem (23). This is feasible by limiting the combined codeblock

[1]The MATLAB scripts and functions for reproducing the results of this paper are available at http://systems.cttc.es/publications/

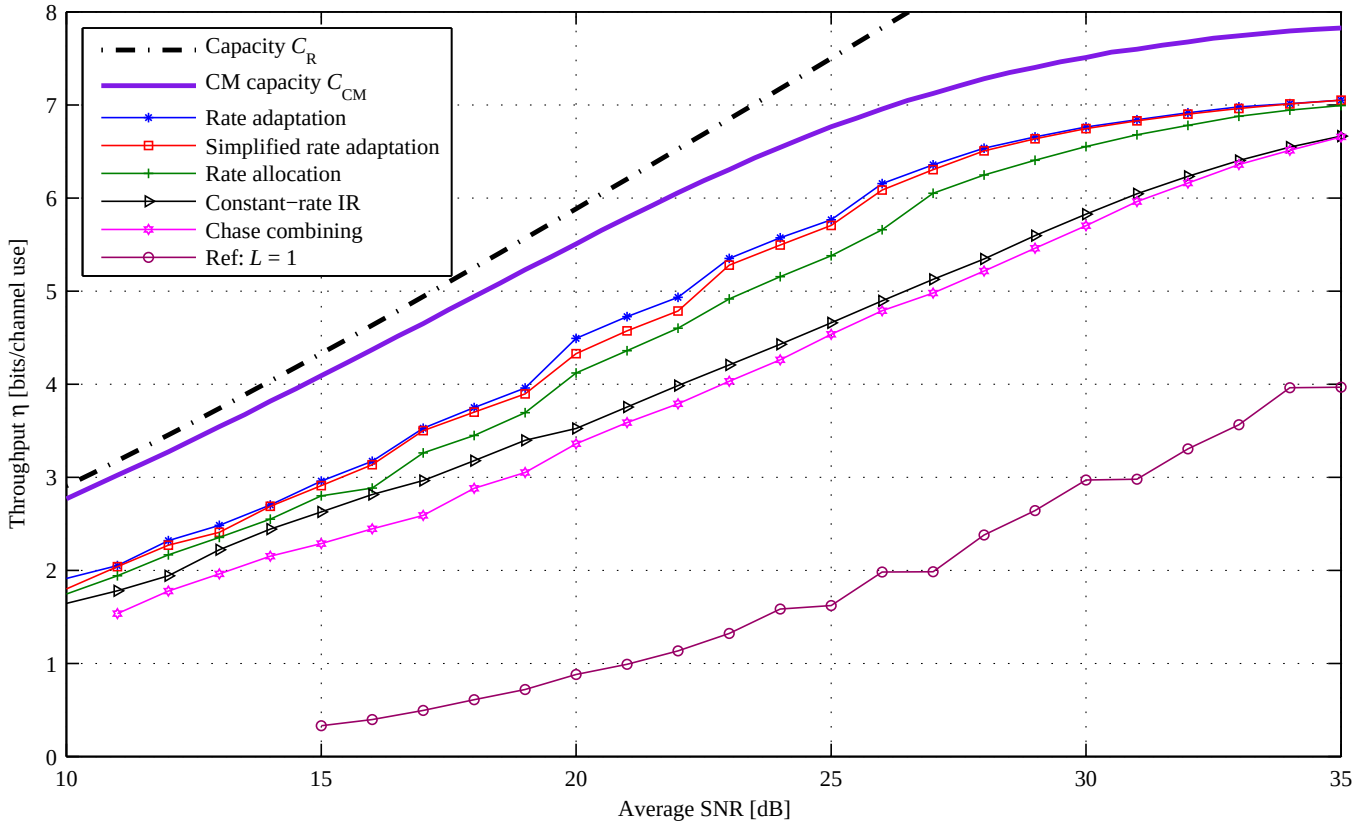


Figure 7. Achieved throughput with various HARQ schemes for a maximum number of $L = 3$ transmissions

lengths to the values given by (19), i.e.

$$\sum_{i=1}^{\ell} N_i \in \mathcal{N}_c,\ T_\ell \in \{1, 2\}\,. \tag{27}$$

For a maximum of three transmissions, we obtain the values given in Table I. It is interesting to observe that there seems to be no simple trend for the blocklengths and modulations as a function of the average SNR $\bar{\gamma}$. This is in line with the findings of Szczecinski et al. [14] who based their analysis on outage probabilities and found a complicated non-convex function for the throughput which could not be solved analytically.

The numerical approach presented here is feasible for a moderate number of retransmissions and code rates and delivers the optimum block lengths and modulations for a given average SNR. These values can be computed offline and stored in a table indexed by the average SNR $\bar{\gamma}$. Note that, aside from numerical approximations, this approach is exact and is not based on idealized assumptions.

In Fig. 7 we can observe that this adaptation of the block lengths results in significant gains with respect to CC or IR with constant block lengths.

### *E. Rate Adaptation: Online Computation of Block Lengths and Modulations*

The offline optimization of the throughput $\eta$ involves averaging over the mutual informations $I_1, \ldots, I_L$ according to (22) and is based on the knowledge of the fading distribution. After a transmission has failed, however, the transmitter has additional knowledge about the previously failed transmissions. Let us assume that the $\ell$-th transmission has failed and that the feedback contains in addition to the ACK/NACK message the mutual informations $I_1, \ldots, I_\ell$ of the previous transmissions (this is often called *multibit NACK* or *intelligent NACK*). These mutual informations can be easily obtained at the receiver since the instantenous SNRs have to be computed by the channel estimator anyway. In addition, the previous blocklengths $N_1, \ldots, N_\ell$ and modulation indices $T_1, \ldots, T_\ell$ are known and can be used to simplify the discrete optimization problem (23). The throughput optimization is then

Table I
OPTIMUM BLOCK LENGTHS AND MODULATIONS FOR $L = 3$ AND $P_{\max} = 0.01$

| $\bar{\gamma}$ | 5 dB | 10 dB | 15 dB | 20 dB | 25 dB | 30 dB | 35 dB |
|---|---|---|---|---|---|---|---|
| $N_1$ | 270 | 135 | 180 | 120 | 108 | 100 | 100 |
| $N_2$ | 450 | 225 | 90 | 60 | 42 | 20 | 20 |
| $N_3$ | 360 | 270 | 90 | 90 | 70 | 60 | 30 |
| $T_1$ | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| $T_2$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $T_3$ | 2 | 1 | 2 | 1 | 1 | 1 | 1 |
| $R$ | 0.9 | 1.7 | 2.8 | 4.2 | 5.4 | 6.5 | 7.0 |
| $f_1$[%] | 44.1 | 50.9 | 51.2 | 63.7 | 42.3 | 23.1 | 8.0 |
| $f_2$[%] | 10.5 | 9.6 | 15.3 | 15.2 | 8.6 | 8.3 | 2.5 |
| $f_3$[%] | 1.0 | 0.85 | 0.6 | 0.82 | 0.32 | 0.32 | 0.48 |

carried out over the smaller set of variables $N_{\ell+1},\dots,N_L$, $T_{\ell+1},\dots,T_L$. This approach, based on the throughput expression (17), is heuristic since for the optimum solution we would have to consider that the lengths $N_\ell$ and modulation indices $T_\ell$ are actually functions of the mutual informations $I_1,\dots,I_{\ell-1}$. For $K\to\infty$, this problem has been solved by dynamic programming [15]. While for $K\to\infty$, the value of the ACMI determines the error event $\mathtt{E}_\ell$, for finite block lengths, it only determines its probability. Hence, the approach based on outage probabilities in the asymptotic regime does not apply directly to the implementation-oriented approach we consider in this paper.

For an ad-hoc solution, we simplify the problem by considering knowledge of decoding failures instead of the full knowledge of the CSI and define the conditional probabilities for $\ell<k\le L$ as

$$f_{k|\ell}\triangleq P\left[\mathtt{E}_{\ell+1}\cdots\mathtt{E}_k\mid\mathtt{E}_1\cdots\mathtt{E}_\ell\right]=\frac{f_k}{f_\ell},\tag{28}$$

which can be computed in analogy to (22) by

$$f_{k|\ell}=\mathbb{E}_{I_{\ell+1}\cdots I_k}\left[\prod_{j=\ell+1}^{k}p_{\mathrm{w}}\left(\sum_{i=1}^{j}I_iN_i,\sum_{i=1}^{j}N_i\right)\right].\tag{29}$$

The throughput is computed with (17) by setting $f_1=\cdots=f_\ell=1$ and replacing $f_{\ell+1},\dots,f_L$ by $f_{\ell+1|\ell},\dots,f_{L|\ell}$. While this is straightforward, for the constraint (23b), we cannot simply replace $f_L$ by its conditional counterpart since this would effectively reduce the target PER. Instead, the constraint becomes

$$f_{L|\ell}<\frac{P_{\max}}{f_\ell}\tag{30}$$

where for $f_\ell$ we can use an estimation based on previous packets or apply the values which are obtained by the offline optimization described above in Subsection IV-D and listed in Table I.

With this approach, the transmitter learns from the outcome of previously failed transmissions. While under the block fading assumption, the transmitter cannot obtain any further knowledge about the channel apart from the average SNR, it can gain knowledge on the amount of the received information in previous transmission attempts. Note that this "learning" only refers to previous transmissions of the same packet and not to an update of a rate adaptation policy in the sense of [47]. After each failed retransmission, the discrete optimization problem (23b) is solved for the parameters $N_{\ell+1},\dots,N_L$, $T_{\ell+1},\dots,,T_L$ and the additional constraint (27).

The obvious drawback of this approach is that for the computation of the blocklength and modulation for the next retransmission, all blocklengths and modulations until the maximum number of transmissions have to be calculated. In other words, if after the $\ell$-th transmission a NACK message is received, the parameters $N_{\ell+1},\dots,N_L$, $T_{\ell+1},\dots,T_L$ are computed via (29) and (17), although only $N_{\ell+1}$ and $T_{\ell+1}$ are required. A method for avoiding this computational overhead is presented in the following.

### *F. Rate Adaptation: Simplified Online Computation*

A simple way to exploit knowledge of $I_1,\dots,I_\ell$ for the calculation of $N_{\ell+1}$, $T_{\ell+1}$ can be obtained via (29), which simplifies to

$$f_{\ell+1|\ell}=\mathbb{E}_{I_{\ell+1}}\left[p_{\mathrm{w}}\left(\sum_{i=1}^{\ell+1}I_iN_i,\sum_{i=1}^{\ell+1}N_i\right)\right]=\frac{f_{\ell+1}}{f_\ell}.\tag{31}$$

Instead of carrying out the complete optimization according to (23), a suboptimum but less complex method is to compute $N_{\ell+1}$, $T_{\ell+1}$ such that (31) is approximately fulfilled while the expected throughput is maximized:

$$\max_{N_{\ell+1},T_{\ell+1}}\ \eta\tag{32a}$$

$$\text{s.t. } f_{\ell+1|\ell}\le\frac{f_{\ell+1}}{f_\ell}.\tag{32b}$$

using the precomputed values for $f_\ell$ according to Subsection IV-D. This method can therefore be seen as a simpler heuristic approximation of the throughput maximization problem. Since only the distribution of the mutual information for the next transmission attempt is required, this method can also exploit partial CSIT. The Appendix outlines a channel model with partial CSIT, which additionally exploits SNR estimates from the previous transmission. Another simple heuristic scheme, which does not consider explicitely a maximum number of retransmissions and is therefore not directly comparable, can be found in [17].

### *G. Simulation Results and Discussion*

Summarizing, we can identify three mechanisms for improving an HARQ scheme based on Chase combining:

1) Incremental Redundancy, for which non-binary repetition coding is particurly well suited.
2) Rate allocation with variable-rate retransmissions. The block lengths and modulation indices can be precomputed if the fading distribution is known.
3) Rate adaptation based on feedback from previously failed transmission.

The first advance alone leads to moderate performance gains without an impact on the system complexity: the LLR-value combining (8) is comparable to maximum-ratio combining. On the other hand, allowing variable block lengths and modulations of the retransmitted blocks increases the performance significantly, in particular for high SNR. Rate allocation precomputes the block lengths and modulations while a further performance improvement is obtained by rate adaptation which computes the block length and modulation based on the channel states of the failed transmissions. The improvements with rate adaptation are remarkable since the two schemes we evaluated are based on simplifications of the optimization problem. Nevertheless, both rate adaptation methods are able to exploit knowledge from previously failed transmissions, at the price of increased complexity and with additional feedback.

Fig. 8 shows the PERs obtained by simulation of the described schemes for a target PER of $P_{\max}=0.01$. Nearly all schemes satisfy the constraint and achieve an error rate

slightly below the limit. In particular, for only one transmission ($L = 1$), the PER obtained by simulation of $10^5$ packets is very close to the target PER. For $L = 3$, the variations of the PER are higher but the constraint (23b) is missed only for some few points and to a small extent. This variation is caused by numerical inaccuracies and the fact that the function $p_{\mathrm{w}}(I_w, N)$, which relates PER and mutual information, is an approximation.

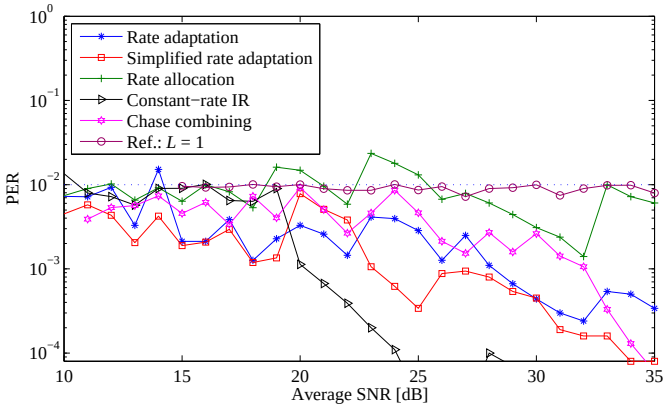


Figure 8. Residual packet error rates after $L$=3 transmissions

## V. Conclusions

We have presented an adaptive HARQ scheme, in which the block lengths and the modulation of the initial transmission and all retransmissions are adapted in order to maximize the throughput while satisfying a constraint on the maximum residual error rate. This optimization can be based either on the channel statistics alone or can additionally consider the channel state of previous transmissions. For the latter approach, two pragmatic schemes have been presented which exploit additional feedback on previously failed transmissions and achieve higher rates than the former rate allocation scheme which is based solely on the channel statistics. We applied a state-of-the-art non-binary LDPC code featuring a repetition scheme which achieves a coding gain without increasing the complexity of the decoder. While this coding scheme is a perfect fit for adaptive HARQ, any channel code for which soft decoding is available can be treated within the same framework.

## Acknowledgment

This work was supported by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM# (Grant agreement no. 318306) and by the Catalan and Spanish Governments under SGR (2009SGR1046) and CICYT (TEC2011-29006-C03-01), respectively. The authors would like to thank the anonymous reviewers, in particular reviewer 1, for their helpful suggestions which helpled to substantially improve the paper.

## Appendix: Partial CSIT

While above we have shown that adaptive HARQ can be successfully appied to the block fading channel, in which the transmitter only knows the average SNR $\bar{\gamma}$, in many situations more knowledge on the channel state might be available. A widely accepted model assumes that the receiver performs channel estimation and prediction and sends the predicted channel coefficients via en error-free feedback link to the transmitter [48]–[51]. The channel coefficients at the transmitter are modelled as

$$\hat{h} = (1-\beta)\, h + v \cdot \sqrt{\beta\,(1-\beta)}, \text{ with } v \sim \mathcal{CN}(0,1), \quad (33)$$

where $\beta \in [0,1)$ is the normalized prediction error, which depends mainly on the average SNR and the user mobility in a wireless system. The unbiased SNR estimate is then

$$\hat{\gamma} = |\hat{h}|^2 \bar{\gamma} + \beta\bar{\gamma}. \quad (34)$$

This leads to the conditional pdf

$$f_{\gamma|\hat{\gamma}}(\gamma \mid \hat{\gamma}) = \frac{1}{\bar{\gamma}\beta} \exp\left(-\frac{\gamma+\hat{\gamma}-\bar{\gamma}\beta}{\bar{\gamma}\beta}\right) \times I_0\left(\frac{2}{\bar{\gamma}\beta}\sqrt{\gamma\,(\hat{\gamma}-\bar{\gamma}\beta)}\right),$$

where $I_0(\cdot)$ denotes the modified Bessel function of the first kind. With the SNR estimate $\hat{\gamma}$, the average SNR $\bar{\gamma}$ and the normalized estimation error $\beta$, the transmitter can apply the online computation of blocklengths described in Section IV-E. To that end, samples with the distribution $f_{\gamma|\hat{\gamma}}$ can be generated by $\gamma = X_1^2 + X_2^2$ where

$$X_1, X_2 \sim \mathcal{N}\left(\sqrt{\frac{\hat{\gamma}-\bar{\gamma}\beta}{2}}, \frac{\bar{\gamma}\beta}{2}\right). \quad (35)$$

From this, samples for the mutual information can be generated and applied to equations (29), (31).

## References

[1] *IEEE Standard for Local and metropolitan area networks. Part 16: Air Interface for Broadband Wireless Access Systems*, IEEE Standard for Local and metropolitan area networks Std., March 2005.

[2] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA);LTE physical layer; General description (3GPP TS 36.201 version 10.0.0 Release 10)*, ETSI Std. TS 136 201 V10.0.0 (2012-03), 2010.

[3] *LTE;Evolved Universal Terrestrial Radio Access (E-UTRA);Medium Access Control (MAC) protocol specification(3GPP TS 36.321 version 10.5.0 Release 10)*, ETSI Std. ETSI TS 136 321 V10.5.0 (2012-03), 2012.

[4] *LTE;Evolved Universal Terrestrial Radio Access (E-UTRA);Physical layer procedures(3GPP TS 36.213 version 10.5.0 Release 10)*, ETSI Std. TS 136 213 V10.5.0 (2012-03), 2010.

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

[6] S. Sesia, G. Caire, and G. Vivier, "Incremental redundancy hybrid ARQ schemes based on low-density parity-check codes," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 1311–1321, Aug. 2004.

[7] R. Negi and J. M. Cioffi, "Delay-constrained capacity with causal feedback," *IEEE Trans. Inform. Theory*, vol. 48, no. 9, pp. 2478–2494, Sept. 2002.

[8] D. V. Djonin, A. K. Karmokar, and V. K. Bhargava, "Joint rate and power adaptation for type-I hybrid ARQ systems over correlated fading channels under different buffer-cost constraints," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 421–435, Jan. 2008.

[9] P. Wu and N. Jindal, "Performance of hybrid-ARQ in block fading channels: A fixed outage probability analysis," *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1129–1141, April 2010.

[10] X. Lagrange, "Throughput of HARQ protocols on a block fading channel," *IEEE Commun. Lett.*, vol. 14, no. 3, pp. 257–259, March 2010.

[11] J. S. Harsini, F. Lahouti, M. Levorato, and M. Zorzi, "Analysis of non-cooperative and cooperative type II hybrid ARQ protocols with AMC over correlated fading channels," *IEEE Trans. Wireless Commun.*, vol. 10, no. 3, pp. 877–889, March 2011.

[12] J.-F. T. Cheng, Y.-P. E. Wang, and S. Parkvall, "Adaptive incremental redundancy," in *IEEE Vehicular Technology Conference (VTC)*, Orlando, FL, USA, Oct. 2003.

[13] G. Yue and X. Wang, "Adaptive hybrid ARQ in Gaussian and turbo coded systems," in *IEEE Globecom*, New Orleans, Nov. 2008.

[14] L. Szczecinski, C. Correa, and L. Ahumada, "Variable-rate transmission for incremental redundancy hybrid ARQ," in *IEEE Globecom*, Miami, FL, USA, Dec. 2010.

[15] L. Szczecinski, S. R. Khosravirad, P. Duhamel, and M. Rahman, "Rate allocation and adaptation for incremental redundancy truncated HARQ," *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2580–2590, June 2013.

[16] T. Villa, R. Knopp, and R. Merz, "Dynamic resource allocation for time-varying channels in next generation cellular networks Part I: A mathematical framework," *IEEE Trans. Wireless Commun.*, 2013, submitted.

[17] S. Pfletschinger and M. Navarro, "Adaptive HARQ for imperfect channel knowledge," in *International ITG Conference on Source and Channel Coding*, Siegen, Germany, Jan. 2010.

[18] J.-F. T. Cheng, "Coding performance of hybrid ARQ schemes," *IEEE Trans. Commun.*, vol. 54, no. 6, pp. 1017–1029, June 2006.

[19] E. Visotsky, Y. Sun, V. Tripathi, M. L. Honig, and R. Peterson, "Reliability-based incremental redundancy with convolutional codes," *IEEE Trans. Commun.*, vol. 53, no. 6, pp. 987–997, June 2005.

[20] K. Kasai, D. Declercq, C. Poulliat, and K. Sakaniwa, "Rate-compatible non-binary LDPC codes concatenated with multiplicative repetition codes," in *IEEE International Symposium on Information Theory (ISIT)*, Austin, TX, USA, June 2010.

[21] C. Poulliat, M. Fossorier, and D. Declercq, "Design of regular $(2,d_c)$-LDPC codes over GF($q$) using their binary images," *IEEE Trans. Commun.*, vol. 56, no. 10, pp. 1626–1635, Oct. 2008.

[22] M. C. Davey and D. MacKay, "Low-density parity check codes over GF($q$)," *IEEE Commun. Lett.*, vol. 2, no. 6, pp. 165–167, June 1998.

[23] J. Huang and J. Zhu, "Linear time encoding of cycle GF($2^p$) codes through graph analysis," *IEEE Commun. Lett.*, vol. 10, no. 5, pp. 369–371, May 2006.

[24] W. Chen, L. Yin, and J. Lu, "Efficient encoding of cycle codes on graphs with large girths," in *IEEE International Conference on Communications, Circuits and Systems (ICCCAS)*, Xiamen, China, May 2008.

[25] A. Bennatan and D. Burshtein, "Design and analysis of nonbinary LDPC codes for arbitrary discrete-memoryless channels," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 549–583, Feb. 2006.

[26] S. Pfletschinger and D. Declercq, "Non-binary coding for vector channels," in *IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, San Francsico, USA, June 2011.

[27] D. Declercq and M. Fossorier, "Decoding algorithms for nonbinary LDPC codes over GF($q$)," *IEEE Trans. Commun.*, vol. 55, no. 4, pp. 633–643, April 2007.

[28] V. Savin, "Min-max decoding for non binary LDPC codes," in *IEEE International Symposium on Information Theory (ISIT)*, Toronto, Canada, July 2008.

[29] E. Boutillon and L. Conde-Canencia, "Bubble check: a simplified algorithm for elementary check-node processing in extended min-sum non-binary LDPC decoders," *IEE Electronic Letters*, vol. 46, no. 9, April 2010.

[30] E. Li, D. Declercq, and K. Gunnam, "Trellis-based extended min-sum algorithm for non-binary LDPC codes and its hardware structure," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2600–2611, July 2013.

[31] J. Hagenauer, E. Offer, and L. Papke, "Iterative decoding of binary block and convolutional codes," *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 429–445, March 1996.

[32] J. Thorpe, "Low-density parity-check (LDPC) codes constructed from protographs," NASA JPL, Tech. Rep. IPN Progress Report 42-154, Aug. 2003.

[33] G. Liva, S. Song, L. Lan, Y. Zhang, S. Lin, and W. Ryan, "Design of LDPC codes: A survey and new results," *Journal of Communication Software and Systems (JCOMSS). Special issue on channel coding in wireless systems*, Sept. 2006.

[34] J. Ha, J. Kim, D. Klinc, and S. W. McLaughlin, "Rate-compatible punctured low-density parity-check codes with short block lengths," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 728–738, 2006.

[35] D. Declercq, S. Pfletschinger, I. Andriyanova, E. Biglieri, G. Boccolini, and C. Poulliat, "D4.8 performance comparison of NB-LDPC, NB-root-LDPC, and NB-DGLDPC," European FP7 project DAVINCI, Tech. Rep., July 2010. [Online]. Available: http://www.ict-davinci-codes.eu/

[36] D. Declercq, V. Savin, and S. Pfletschinger, "Multi-relay cooperative NB-LDPC coding with non-binary repetition codes," in *The Eleventh International Conference on Networks ICN*, Saint Gilles, Reunion Island, Feb. 2012.

[37] K. Kasai, D. Declercq, and K. Sakaniwa, "Fountain coding via multiplicatively repeated non-binary LDPC codes," *IEEE Trans. Commun.*, vol. 60, no. 8, pp. 2077–2083, Aug. 2012.

[38] S. Pfletschinger, A. Mourad, E. López, D. Declercq, and G. Bacci, "Performance evaluation of non-binary LDPC codes," in *ICT-MobileSummit*, Santander, Spain, June 2009.

[39] N. J. A. Sloane, "Tables of sphere packings and spherical codes," *IEEE Trans. Inform. Theory*, vol. 27, no. 3, pp. 327–338, May 1981.

[40] J. Hagenauer, "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," *IEEE Trans. Commun.*, vol. 36, no. 4, pp. 389–400, April 1988.

[41] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 927–946, May 1998.

[42] M. Franceschini, G. Ferrari, and R. Raheli, "Does the performance of LDPC codes depend on the channel?" *IEEE Trans. Commun.*, vol. 54, no. 12, pp. 2129–2132, Dec. 2006.

[43] K. Brüninghaus, D. Astély, T. Sälzer, S. Visuri, A. Alexiou, S. Karger, and G.-A. Seraji, "Link performance models for system level simulations of broadband radio access systems," in *IEEE PIMRC*, Berlin, Germany, Sept. 2005.

[44] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[45] S. Pfletschinger and M. Navarro, "Link adaptation with retransmissions for partial channel state information," in *IEEE Globecom*, New Orleans, USA, Nov. 2008.

[46] S. H. Kim, S. J. Lee, and D. K. Sung, "Low-complexity rate selection of HARQ with Chase combining in Rayleigh block-fading channels," *IEEE Trans. Veh. Technol.*, vol. 62, no. 6, pp. 2818–2824, July 2013.

[47] S. R. Khosravirad, L. Szczecinski, and F. Labeau, "Rate-adaptive HARQ in relay-based cooperative transmission," in *IEEE International Conference on Communication (ICC)*, Budapest, Hungary, June 2013.

[48] T. Ekman, "Prediction of mobile radio channels: Modeling and design," Ph.D. dissertation, Uppsala University, Sept. 2002. [Online]. Available: http://www.signal.uu.se/Publications/ptheses.html

[49] T. Ekman, M. Sternad, and A. Ahlén, "Unbiased power prediciton of Rayleigh fading channels," in *IEEE VTC*, Vancouver, Canada, Sept. 2002.

[50] S. Falahati, A. Svensson, T. Ekman, and M. Sternad, "Adaptive modulation systems for predicted wireless channels," *IEEE Trans. Commun.*, vol. 52, no. 2, pp. 307–316, Feb. 2004.

[51] A. Duel-Hallen, "Fading channel prediction for mobile radio adaptive transmission systems," *Proc. IEEE*, vol. 95, no. 12, pp. 2299–2313, Dec. 2007.