Joint Coding/Decoding for Multi-message HARQ

Abdellatif Benyouss, Mohammed Jabi, Maël Le Treust, Leszek Szczecinski

Joint Coding/Decoding for Multi-message HARQ

Abdellatif Benyouss, Mohammed Jabi, Maël Le Treust* and Leszek Szczecinski
INRS-EMT, University of Quebec, Montreal, Canada.
e-mail: {benyouss,jabi,leszek}@emt.inrs.ca
*ETIS - UMR 8051 / ENSEA - University Cergy-Pontoise - CNRS, France.
e-mail: mael.le- treust@ensea.fr.

***Abstract*—In this work, we propose and investigate a new coding strategy devised to increase the throughput of hybrid ARQ (HARQ) transmission over block fading channel. In our proposition, the transmitter jointly encodes a variable number of bits for each round of HARQ. The parameters (rates) of this joint coding can vary and may be based on the negative acknowledgment (NACK) signals provided by the receiver or, on the past (outdated) information about the channel states. The results indicate that significant gains can be obtained using the proposed coding strategy especially where the conventional HARQ fails to offer throughput improvement even if the number of transmission rounds is increased.**

I. Introduction

In this work we propose a new coding strategy to increase the throughput of HARQ transmission over block-fading channel. Our approach consists in a joint coding/decoding of multiple messages whose contents depend on the outcome of the decoding in the HARQ transmission rounds.

HARQ is commonly used in modern communications systems to deal with unpredictable changes in the channel– e.g., due to fading, and with the distortion of the transmitted signals–due to noise. HARQ relies on the feedback channel over which the receiver informs the transmitter about the decoding errors (via NACK messages). After each NACK, the transmitter makes another transmission *round* which conveys additional information necessary to decode the message. This continues till the positive acknowledgment (ACK) message is received and then the HARQ *cycle* starts again for another message. In the so-called *truncated* HARQ, the cycle stops if the maximum number of rounds is attained.

As in many previous works, e.g., [1], we will consider throughput as performance measure, since it represents what the end-users values foremost [2]. In the case of transmission without channel state information (CSI), which we also treat in this work, it was shown in [1] that incremental redundancy HARQ (IR-HARQ) may approach the ergodic capacity of the channel. It requires, however, using a high (infinite, in theory) transmission rate per channel block and high (infinite) number of transmission rounds.

Such a solution has, therefore, a limited practical value since using long buffers becomes the limiting factor for implementation of HARQ [3]. On the other hand, using finite rates and truncated HARQ, the difference between the throughput achievable using HARQ and the theoretical limits may be large especially if we target the spectral efficiencies close to the initial (nominal) transmission rate [4].

To address this problem, various adaptive versions of HARQ were proposed in the literature. Targeting the increase of the throughput. In particular [5]–[11] suggested to vary the duration of the retransmitted codewords to convey enough information to decode the message and yet minimize the number of channel uses. The obvious drawback is that the resources assigned to the various HARQ rounds are not constant. To deal with this issue it is possible to encode many messages into predefined size blocks as done in [12] or to group variable-length packets into long frames [11].

In this work we follow a different path. Namely, we propose to increase the number of information bits encoded in each transmission round. This may be seen as encoding of various messages into a single channel block and we want to adapt the coding rate of each message so that the throughput is maximized. Joint coding of many messages for HARQ was proposed before, e.g., in [13], [14] which targeted, however, the increased transmission reliability for a maximum of two transmissions, and in [15], where two messages were encoded, each with the same rate. We discovered that a similar idea of joint multi-message coding was also proposed in [16] and [17], whose main differences with the current work are i) a lack of formal optimization of the transmission parameters, ii) a simplified, layered decoding scheme in [16], and iii) analysis of decoding errors with short blocks in [17].

The contributions of this work are thus the following:

- We propose the general framework to analyze joint encoding of multiple messages which allows us to derive the relationship between the coding rates and the throughput. Our approach is different from [13] which also considered multi-message coding but focused on increasing of the transmission reliability, and limited considerations to two transmissions. Our proposed coding scheme can be considered as a generalization of [15]–[17], which used predefined type of adaptation of the coding rates to the outdated CSI.
- We propose to use the so-called multi-bit feedback to adapt the transmission rates to the channel state experienced by the receiver in the past transmission rounds of HARQ. This is similar to the idea exploited already in [8], [11], [16], [18]; it simplifies the optimization of the rates and yields the results which may be treated as performance limits of a conventional, one-bit feedback where only the ACK/NACK message is transmitted through the feedback channel.

- We optimize the coding rates using the Markov decision process (MDP) formulation, and compare the proposed, multi-message HARQ to the conventional one from the throughput point of view.

The work is organized as follows: we define the transmission model as well as the basic performance metrics in Sec. II, and explain the idea of multi-message coding in Sec. III. The optimization of the rates in the proposed coding strategy is explained in Sec. IV. The numerical results are presented in form of short examples throughout the work to guide the reader and illustrate the main idea. Conclusions are presented in Sec. V.

II. Channel model and HARQ

Consider the problem of transmitting the message $\mathsf{m} \in \{0,1\}^{RN_\mathrm{s}}$ where R denotes the coding rate per block and N_s is the number of channel uses in a block i.e., RN_s is the number of bits transmitted over each block. The message is first encoded using coding function $\Phi[\cdot]$ into the codeword $\boldsymbol{x} = \Phi[\mathsf{m}] \in \mathcal{X}^{KN_\mathrm{s}}$ composed of KN_s complex symbols taken from the constellation $\mathcal{X}$. Next the codeword $\boldsymbol{x}$ is divided into K subcodewords $\boldsymbol{x}_k, k = 1, \ldots, K$ and each subcodeword is transmitted over a channel *block* whose output is given by

$$\boldsymbol{y}[n] = \sqrt{\mathsf{snr}[n]}\boldsymbol{x}[n] + \boldsymbol{z}[n], \tag{1}$$

where n is the index of the channel block, $\boldsymbol{z}[n]$ is additive white Gaussian noise (AWGN) complex noise with unitary variance, $\sqrt{\mathsf{snr}[n]}$ is the gain of the channel at block n. We model elements of $\boldsymbol{x}$ as independent, identically distributed (i.i.d.) random variables X with $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$, so $\mathsf{snr}[n]$ has the meaning of the signal-to-noise ratio (SNR) at the receiver. The channel gains $\mathsf{snr}[n]$ are random and we model $\mathsf{snr}[n]$ as i.i.d. random variables with distribution $p_{\mathsf{SNR}}(\mathsf{snr})$.

The important elements of this block-fading model are i) in each block the resources used (number of symbols, transmitted power, or bandwidth) are the same, ii) the CSI, $\mathsf{snr}[n]$, although may be perfectly estimated at the receiver, is not available at the transmitter at block-time n, and iii) there is a one-block delayed feedback channel, over which the receiver may inform the transmitter at time $n+1$ whether the decoding of the data sent at time n succeeded (through an ACK) or failed (though a NACK).[1]

We further assume that the transmitter has saturated buffer and delivering it efficiently (i.e., in the shortest time) is formalized as maximization of the throughput, defined as a ratio between the number of correctly received bits and the number of channel used. It may be calculated as follows

$$\eta \triangleq \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathcal{R}[n], \tag{2}$$

where $\mathcal{R}[n]$ is the instantaneous reward, i.e., $\mathcal{R}[n]N_\mathrm{s}$ is the number of correctly received bits after the transmission in the time n. Of course, if the receiver fails to decode the message, we set $\mathcal{R}[n] = 0$.

Since, in general, the message m is transmitted using a varying (possibly in a random manner) number of channel blocks, to calculate the throughput, it is often convenient to replace the time-wise consideration of (2) with a message-wise analysis, which transforms (1) into

$$\boldsymbol{y}_k = \sqrt{\mathsf{snr}_k}\boldsymbol{x}_k + \boldsymbol{z}_k, \quad k = 1, \ldots, K, \tag{3}$$

where k indicates the HARQ round related to the same message, and, in general, the maximum number of transmission rounds of the same packet may be limited to $K < \infty$, i.e., the retransmissions are aborted after the event NACK_K.[2]

Then, to calculate (2) we may use the well known renewal-reward theorem [1] which yields

$$\eta = \frac{\mathbb{E}[\mathcal{R}]}{\mathbb{E}[\mathcal{K}]}, \tag{4}$$

where $\mathcal{R}$ is the reward per message, $\mathcal{K}$ is the number of channel blocks used to transmit the message, and the expectations are taken with respect to all (random) SNRs affecting these two random variables.

For the moment we did not suppose anything about the encoding/decoding operation but, to make the discussion simple, we assume that the mutual information (MI), $I_k = \mathsf{I}(X_k; Y_k|\mathsf{snr}_k)$ between the random variables X_k and Y_k, modelling, respectively, the channel input and output in the kth round is a sufficient measure to determine the decoding success or failure under assumption of the random coding and maximum likelihood (ML) decoding. Of course, the MI I_k depends on snr_k, i.e., $I_k = I(\mathsf{snr}_k)$.

Under assumption of the CSI not being available at the transmitter, the highest achievable throughput is given by the ergodic capacity[3] of the channel

$$\overline{C} \triangleq \mathbb{E}_{\mathsf{SNR}}[I(\mathsf{SNR})] = \mathsf{I}(X_k; Y_k|\mathsf{SNR}). \tag{5}$$

The "straightforward" coding strategy which achieves $\overline{C}$ may be done as follows: first encode a message $\mathsf{m} \in \{0,1\}^{K\overline{C}N_\mathrm{s}}$ into a codeword of KN_s symbols, then divide the codeword into K disjoint subcodewords $\boldsymbol{x}_k = \Phi_k[\mathsf{m}], k = 1, \ldots, K$ each composed of N_s symbols, and finally we transmit one subcodeword after another over K blocks. The observations $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_K$ are stored at the receiver, i.e., the MI is accumulated over the transmissions blocks and thus, the probability of decoding errors goes asymptotically to zero, i.e., $\lim_{K\to\infty} \Pr\left\{\sum_{k=1}^{K} I_k < K\overline{C}\right\} = 0$. A large number of blocks K may be necessary to approach closely the capacity and it requires the receiver to store the signals $\boldsymbol{y}_k, k = 1, \ldots, K$, which is not practical because memory limitations are a constraint in the design of the receivers [3].

[1] We also assume that the errors detection is possible (e.g., via cyclic redundancy check (CRC) mechanisms) and that the feedback channel is error-free (as can be guaranteed by a strong protection of the one-bit messages). Any loss of resources due to the overhead can be or– taken into account into the rate calculation, or– ignored, if the payload N_s is large.

[2] In such a case, depending on the application requirements, the packet whose decoding failed may be kept in the transmitter's buffer and transmitted again, or it may be discarded.

[3] We use the term "capacity" to denote the achievable rate for a *given* distribution of X.

A. HARQ

Assume now that, as before, we encode the message $\mathsf{m} \in \{0,1\}^{RN_s}$ into subcodewords $\boldsymbol{x}_k = \Phi_k[\mathsf{m}], k = 1,2,\ldots,$ each transmitted over the kth channel block. But, unlike in the "straightforward" strategy aiming to attain the ergodic capacity, the presence of the feedback will be exploited and the transmission cycle will abort when the decoding is successful (as confirmed by an ACK feedback message). In this case the number of blocks used to successfully deliver the message is a random variable $\mathcal{K}$; to calculate (4) we need to find the expectation of $\mathcal{K}$.

Let us denote by NACK_k the event of decoding failure after k rounds of the same packet

$$\mathsf{NACK}_k = \left\{ \left(I_1 < R\right) \wedge \left(I_2^{\Sigma} < R\right) \wedge \ldots \wedge \left(I_k^{\Sigma} < R\right) \right\} \quad (6)$$

$$= \left\{ I_k^{\Sigma} < R \right\}, \quad (7)$$

where $I_k^{\Sigma} \triangleq \sum_{t=1}^{k} I_t$ is the MI accumulated throughout the k rounds; relating the decoding success/failure to accumulated MI means that the codewords $\boldsymbol{x}_k$ are generated independently which corresponds to IR-HARQ.

The probability of successful decoding in the kth transmission round (and thus, obtaining the reward R) may be calculated as $f_{k-1} - f_k$, where $f_k \triangleq \Pr\{\mathsf{NACK}_k\}, k \geq 1$ and $f_0 \triangleq 1$ [1]. Then (4) is calculated as follows

$$\eta_K = \frac{R(1-f_1) + R(f_1 - f_2) + \ldots + R(f_{K-1} - f_K)}{1\cdot(1-f_1) + 2\cdot(f_1 - f_2) + \ldots + K\cdot(f_{K-1})} \quad (8)$$

$$= \frac{R(1-f_K)}{1 + \sum_{k=1}^{K-1} f_k}. \quad (9)$$

Example 1 (Two-states channel)**.** *Consider a block-fading channel where the channel can only takes two values of MI: I_a and $I_b > I_a$ with $\Pr\{I = I_a\} = 1-p$ and $\Pr\{I = I_b\} = p$, so the ergodic capacity is given by $\overline{C} = I_a(1-p) + I_b p$. Assuming $I_a > 0$ we may also force HARQ to deliver the message at most in the last transmission, i.e., to ensure that $f_K = 0$, which means that we impose the constraints on the transmission rate $R \leq K I_a$.*

Assume $I_a = 1, I_b = 1.5$, and $p = 0.75$ so $\overline{C} = 1.375$. For $K = 2,3$ we easily calculate the throughput[4]

$$\eta_2 = \begin{cases} R & \text{if} \quad R \leq 1 \\ 0.8R & \text{if} \quad 1 < R \leq 1.5 \\ 0.5R & \text{if} \quad 1.5 < R \leq 2 \end{cases} \quad (10)$$

$$\eta_3 = \begin{cases} \eta_2 & \text{if} \quad R \leq 2 \\ 0.48R & \text{if} \quad 2 < R \leq 2.5 \\ 0.41R & \text{if} \quad 2.5 < R \leq 3 \end{cases} \quad (11)$$

The optimum throughput-rate pairs are then $(R = 1.5, \eta_2 = 1.2)$ and $(R = 3, \eta_3 = 1.23)$. We first note that the benefit of using HARQ is clear: we are able to transmit without errors with a finite number of channel blocks and go beyond the obvious limit of I_a. Second, we note that for $K = 2$, after two transmissions, the accumulated MI always satisfies $I_2^{\Sigma} \geq 2$, while the condition $I_2^{\Sigma} \geq 1.5$ is sufficient to decode the message. This may seem as the "waste" which we will remove with the idea of multi-message coding introduced in Sec. III.

[4]For $R \leq 1$ we have $f_1 = 0$. For $1 < R \leq 1.5$, $f_1 = 1-p$ and $f_2 = 0$. For $1.5 < R \leq 2$, $f_1 = 1$, $f_2 = 0$, etc.

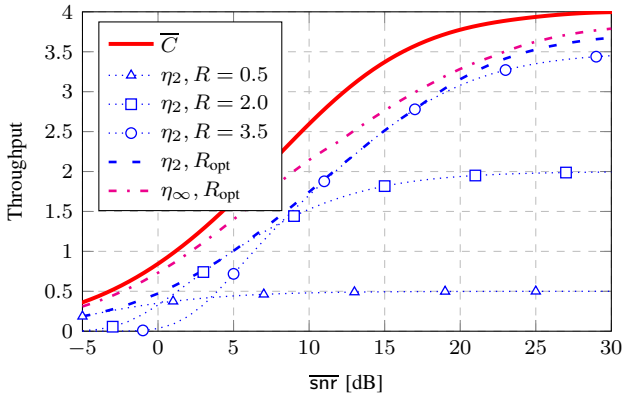

Fig. 1. Throughput of the conventional IR-HARQ, compared to the ergodic capacity, $\overline{C}$, in Rayleigh block-fading channel. The R_{opt} curve is an envelope of the throughputs η_K obtained with different transmission rates per block $R \in \{0.25, 0.5, \ldots, 7.75\}$.

Example 2 (16-QAM over Rayleigh fading channel)**.** *Assume now that the transmission is done using symbols drawn uniformly from 16-points quadrature amplitude modulation (QAM) constellation $\mathcal{X}$ [19, Ch. 2.5] and that the channel gains follow Rayleigh distribution, i.e.,*

$$p_{\mathsf{SNR}}(\mathsf{snr}) = 1/\overline{\mathsf{snr}} \exp(-\mathsf{snr}/\overline{\mathsf{snr}}), \quad (12)$$

where $\overline{\mathsf{snr}}$ is the average SNR.

We calculate $I(\mathsf{snr})$ and $\overline{C}$ numerically as shown in [19, Ch. 4.5] and compare it in Fig. 1 with the throughput η_∞. The results indicate that i) there is a significant loss with respect to the ergodic capacity when using truncated HARQ (here $K = 2$) and ii) increasing the number of transmission rounds ($K = \infty$) helps recovering the loss for a small-medium range of throughput (e.g., for $\eta_\infty = 1$ we gain ~ 3 dB comparing to $\eta_2 = 1$ so the gap to $\overline{C}$ is less than 1 dB), but it less useful in the region of high throughput, i.e., in the vicinity of the maximum attainable transmission rate (e.g., for $\eta_\infty = 3$, we gain 1dB with respect to $\eta_2 = 3$ but the gap to $\overline{C}$ is still ~ 5 dB). We highlight this known effect [4] to emphasize later the gains of the new coding strategy.

III. Multi-message HARQ

The examples we have shown previously indicate that the conventional coding cannot bring the throughput of HARQ close to the capacity unless we start to increase the coding rate R and the number of rounds K. We would like now to exploit the new coding possibility and add messages during the HARQ cycle.

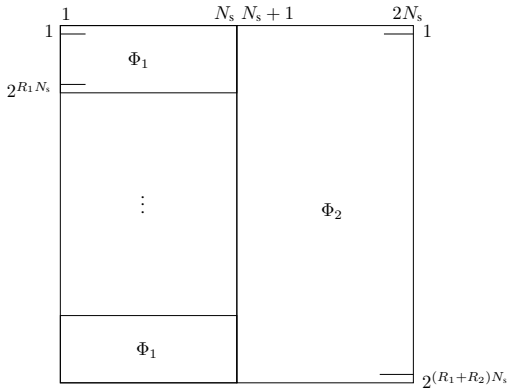

Fig. 2. Illustration of the codebook defined through the coding function Φ_1 in (13) and the joint coding function Φ_2 in (14). Each codeword composed of $2N_s$ symbols is indexed by the message m. The first N_s symbols are created without indexing by m_2 so we artificially repeat them $2^{R_2 N_s}$ times to match the number of codewords in the codebook Φ_2.

Let us start with the case of two transmission rounds. In the first round, we use the rate R_1, i.e., the message $\mathsf{m}_1 \in \{0,1\}^{R_1 N_s}$ is encoded

$$\boldsymbol{x}_1 = \Phi_1[\mathsf{m}_1] \in \mathcal{X}^{N_s}, \tag{13}$$

transmitted over the channel (1) yielding the outcome $\boldsymbol{y}_1 = \sqrt{\mathsf{snr}_1}\boldsymbol{x}_1 + \boldsymbol{z}_1$.

After the decoding failure (which occurs if $I_1 < R_1$) we encode the message $\mathsf{m} = [\mathsf{m}_1, \mathsf{m}_2] \in \{0,1\}^{R_2^{\Sigma} N_s}$ using a conventional code designed independently of the codebook corresponding to the first transmission

$$\boldsymbol{x}_2 = \Phi_2[\mathsf{m}] \in \mathcal{X}^{N_s}, \tag{14}$$

which yields the channel outcome $\boldsymbol{y}_2 = \sqrt{\mathsf{snr}_2}\boldsymbol{x}_2 + \boldsymbol{z}_2$. Here, $R_2^{\Sigma} = R_1 + R_2$ where R_2 can be seen as the nominal coding rate of the message m_2. This coding strategy is introduced without any claim to optimality but with the obvious advantage of being simple to analyze. The idea of using Φ_2 independent of Φ_1 was also proposed in [17].

Intuitively, by introducing m_2 we want to prevent the "waste" of MI, which happens if I_2^{Σ} is much larger than R_1, cf. Example 1.

After the second transmission, the decoder wants to recover the messages $\mathsf{m}_1, \mathsf{m}_2$ using the observations $\boldsymbol{y}_1, \boldsymbol{y}_2$. The codebook obtained in two transmissions is illustrated in Fig. 2. The associated decoding conditions after in the second transmission round are given by

$$\mathsf{I}(X_1;Y_1) + \mathsf{I}(X_2;Y_2) > R_1 + R_2, \tag{15}$$

$$\mathsf{I}(X_2;Y_2) > R_2, \tag{16}$$

where (15) is a constraint over the sum-rate that guarantees the joint decoding of the pair of messages $(\mathsf{m}_1, \mathsf{m}_2)$ and (16) corresponds to the correct decoding of the message m_2. Intuitively, the MI must be accumulated to decode each of the messages even though the decoding is done jointly; the formal proof of (15) and (16) follows conventional information-theoretic arguments; we skip it for brevity though. We note that the similar idea of increasing the rates of retransmission appeared already in [20] in the context of physical layer (PHY) security.

While the event of NACK_1 remains unchanged with respect to the conventional coding, the event NACK_2 has a different meaning. Namely, using negation of the conditions shown in (15) and (16) we obtain

$$\begin{aligned}\mathsf{NACK}_2 &= \Big\{\big(I_1 < R_1\big) \wedge \overline{\big((I_2^{\Sigma} > R_1 + R_2) \wedge (I_2 > R_2)\big)}\Big\} \\ &= \Big\{\big(I_1 < R_1\big) \wedge \big((I_2^{\Sigma} < R_1 + R_2) \vee (I_2 < R_2)\big)\Big\} \\ &= \Big\{\big(I_1 < R_1\big) \wedge \big(I_2^{\Sigma} < R_1 + R_2\big)\Big\},\end{aligned} \tag{17}$$

where $\overline{A}$ is a negation of A. Here, we find that (16) is weaker than all other conditions and does not appear in (17).

For more than two rounds, the above conditions generalize quite straightforwardly as

$$\begin{aligned}\mathsf{NACK}_k = \Big\{\big(I_1 < R_1\big) \wedge \big(I_2^{\Sigma} < R_2^{\Sigma}\big) \wedge \\ \ldots \wedge \big(I_k^{\Sigma} < R_k^{\Sigma}\big)\Big\},\end{aligned} \tag{18}$$

where $R_k^{\Sigma} \triangleq \sum_{t=1}^{k} R_t$, and R_k is the rate of the message m_k added in the kth round.

To calculate the throughput of the multi-message (mm) HARQ, we adopt the approach similar to the one used in (8) but we must account for the reward in the k transmission round given by R_k^{Σ}. This yields

$$\begin{aligned}\eta_K^{\mathrm{mm}} &= \frac{R_1^{\Sigma}(1-f_1) + R_2^{\Sigma}(f_1 - f_2) + \ldots + R_K^{\Sigma}(f_{K-1} - f_K)}{(1-f_1) + 2\cdot(f_1 - f_2) + \ldots + K\cdot(f_{K-1})} \\ &= \frac{\sum_{k=1}^{K} R_k\big(f_{k-1} - f_K\big)}{\sum_{k=0}^{K-1} f_k}.\end{aligned} \tag{19}$$

The fundamental difference of the proposed coding strategy with respect to those available in the literature appears now clearly in the numerator of (19) which expresses the idea of variable rate transmission (due to encoding of multiple messages). Nevertheless, not only the numerator changed with respect to (9) but also the denominator is different due to the new definition of NACK_k in (18).

Also, if we set $R_k = 0, k = 2, \ldots, K$, (19) is equivalent to (9); we thus recover the conventional, single-message HARQ.

Example 3 (Two-state channel and multi-message coding)**.** *We can consider now the proposed multi-message joint coding and decoding in the scenario of Example 1, where for a fair comparison of the results obtained we will force the successful decoding at the final round, i.e., $f_K = 0$.*

Let us start, as before, with $R_1 = 1.5$. In the case of decoding failure (which means that we obtained $I_1 = I_a = 1$), we are free to define the rate R_2 as we wish but, in the absence of any formal criterion (more on that in Sec. IV-B) we take the auxiliary (and somewhat ad-hoc) condition: namely, we want to guarantee a non-zero decoding probability, i.e., $f_2 < 1$. Here, since $I_2^{\Sigma} \in (2, 2.5)$, we can use any $R_2 \leq 1$ which ensures $f_2 < 1$; on the other hand, using the rate $R_2 \leq 0.5$ we guarantee a much stronger condition $f_2 = 0$.

First $K=2$; using $R_2=0.50$, we obtain $f_1=0.25$ and $f_2=0$ so the throughput is given by

$$\eta_2^{\mathrm{mm}}=\frac{R_1+0.25R_2}{1+0.25}=1.30, \tag{20}$$

where we note that we used exactly the same channel resources as in the conventional HARQ and we obtained the same guarantee of successful decoding ($f_2=0$) after two transmission rounds.

The difference is that, while we still have $I_2^{\Sigma}\in(2,2.5)$, we now use $R_2^{\Sigma}=2$ and thus it is necessary to have $I_2^{\Sigma}=2$ in order to decode both messages; in a way we eliminated the waste of MI in the conventional IR-HARQ, where $R^{\Sigma}=1.5$. The improvement may be seen as the increase in the throughput (from $\eta_2=1.20$ to $\eta_2^{\mathrm{mm}}=1.30$) or as the reduction in the memory requirements (i.e., we obtain a better throughput with smaller K, see $\eta_3=1.23$ in Example 1). This does not come for free: we will have to pay with the possibly increased complexity of multi-message joint encoding/decoding.

For $K=3$ we can use the larger value of R_2 (that guarantees $f_2<1$), i.e., $R_2=1$. In this case, $f_1=0.25$, and $f_2=\Pr\left\{I_1<1.50\wedge I_2^{\Sigma}<2.5\right\}=0.0625$. In the third transmission we observe $I_3^{\Sigma}\in(3,3.5)$ so, using $R_3=0.5$, we obtain $f_3=0$ and thus the throughput is calculated as

$$\eta_3^{\mathrm{mm}}=\frac{R_1+0.25R_2+0.0625R_3}{1+0.25+0.0625}\approx 1.36, \tag{21}$$

which is already quite close to $\overline{C}=1.375$.

IV. Performance limits and optimization of the coding rates

The objective now is to evaluate how well the proposed multi-message IR-HARQ can perform and, to this end, we will have to find the optimal coding rates $R_1,R_2,\ldots,R_K$ which maximize throughput (19). This optimization problem is quite difficult and we will thus proceed in two steps: we will first assume that the value of the accumulated MI can be transmitted over the feedback channel and next, we will optimize the rates using this additional information.

A. Multi-bit feedback

The two-states channel shown in Example 3, was quite simple and each NACK message not only informed us about the necessity of subsequent rounds, but also, provided us with a valuable information about the state of the decoder. Indeed, after the kth round we knew exactly the value of I_k^{Σ} (because only two values of MI were possible). Then, for a given I_k^{Σ} the only random element in the $(k+1)$th round is the MI, I_{k+1}, and this greatly simplified the choice of the rate R_{k+1}.

We will take this idea further, and assume that, after kth round, the receiver sends over the feedback channel the value of I_k^{Σ}. This requires the feedback channel to support multi-bits messages on top of one-bit acknowledgements. For simplicity, we will neglect the overhead due to this additional signalling; this is justified when N_{s} is large.

We emphasize that I_k^{Σ} contains the obsolete CSI and we do not violate assumptions made in Sec. II because I_k^{Σ}, being available at the transmitter at $(k+1)$st round, cannot be used to infer anything about I_k (due to i.i.d. model of the SNRs). In fact, adaptation of the transmission parameters in HARQ on the basis of the obsolete CSI was already considered before, e.g., in [5], [7]–[11]. Of course, such a feedback cannot be used in the conventional IR-HARQ because, the transmission rate, R, is the only parameter of the HARQ and is determined before HARQ starts.

Beside the advantage of simplifying the optimization, this approach will also tell us what are the limits of HARQ protocol with any (obsolete) feedback.

B. Optimization via MDP

We may now formalize the multi-message HARQ as a MDP where, at time n, being in the *state* $\mathsf{s}[n]\in\mathcal{S}$, we have to take the action $\mathsf{a}[n]=\pi\big(\mathsf{s}[n]\big)\in\mathcal{A}$, where the *policy* π is a mapping between the space of states, $\mathcal{S}$, and the actions, $\mathcal{A}$, i.e., $\pi:\mathcal{S}\mapsto\mathcal{A}$. The actions are the transmission rates, R, and the state must be defined so that the probability of passing from the state $\mathsf{s}[n]=\mathsf{s}'$ to $\mathsf{s}[n+1]=\mathsf{s}''$ depends on the action $\mathsf{a}[n]=R[n]$ and the channel's MI $I[n]$.

The original definition of the throughput we made in (2) is more useful than the one in (19) and we reformulate it as

$$\eta_{\infty}^{\mathrm{mm}}=\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\big[\mathcal{R}(\mathsf{s}[n],\mathsf{a}[n])\big], \tag{22}$$

where the expectations are taken with respect to the random states $\mathsf{s}[n]$, and $\mathcal{R}(\mathsf{s}[n],\mathsf{a}[n])$ is the average reward obtained when taking the action $\mathsf{a}[n]$ in the state $\mathsf{s}[n]$.

We thus define the state as a pair

$$\mathsf{s}[n]\triangleq(I^{\Sigma}[n],R^{\Sigma}[n]), \tag{23}$$

where $R^{\Sigma}[n]$ and $I^{\Sigma}[n]$ are, respectively, accumulated rate and accumulated MI after the transmission in block $n-1$, thus the state at time $n+1$ becomes

$$\mathsf{s}[n+1]=\begin{cases}\big(I^{\Sigma}[n]+I[n],R^{\Sigma}[n]+R[n]\big)\\ \qquad\text{if}\quad R^{\Sigma}[n]+R[n]\geq I^{\Sigma}[n]+I[n]\\ (0,0)\quad\text{otherwise}\end{cases}.$$

Since a new cycle starts only when the messages are correctly decoded, a non-zero reward is obtained only by terminating the HARQ cycle (moving to the state $\mathsf{s}[n+1]=(0,0)$); thus the decoding is necessary to obtain a reward, i.e.,

$$\mathcal{R}(\mathsf{s}[n],R[n])=\big(R^{\Sigma}[n]+R[n]\big)F_I^{\mathrm{c}}(R^{\Sigma}[n]-I^{\Sigma}[n]+R[n]),$$

where $F_I^{\mathrm{c}}(x)\triangleq 1-F_I(x)$ and $F_I(x)$ is the cumulative density function (CDF) of I.

Optimization of (22) with respect to the policy π may be then done very efficiently using known algorithms such as a *policy iteration* [21, Ch. 2]. For brevity we omit a detailed explanation of the algorithm but we follow closely the approach used in [15], [18].

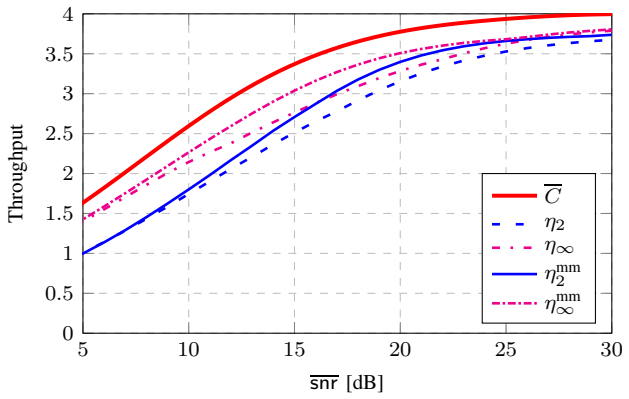

Fig. 3. Rayleigh block-fading channel: throughput, η_K^{mm}, of the proposed multi-message IR-HARQ (optimized over transmission rates) is compared to the throughput of the conventional IR-HARQ, $\eta_{\underline{K}}$, (also optimized over transmission rates), and to the ergodic capacity, $\overline{C}$. The range of SNR is changed with respect to Fig. 1 to clearly show the region where the improvement is notable.

We emphasize here that while assuming the feedback about the outdated MI is discretized with an arbitrary resolution when optimizing the throughput, this does not mean necessarily that number of bits in the feedback channel must be also increased to achieve the same performance. In fact, since in practice only a limited number of rates is available, obviously the same rate will be attributed to many different values of the MI. Thus, the receiver can compute the accumulated MI and only transmits the index of the optimal rate which will be available in a table computed off-line.

Example 4 (16-QAM over Rayleigh fading channel (continued))**.** *In order to optimize the throughput* (22) *we discretize the space of states using the approach outlined already in [18]. Namely, the MI, which is the first variable defining the state in* (23)*, is discretized over* 2^6 *points and we choose the actions* R *to belong to the set* $\{0.25, 0.5, 0.75, \ldots, 7.75\}$*, which also imposes the same discretization of the second variable,* R^{Σ}*, defining the state in* (23)*.*

The results obtained are shown in Fig. 3, where the improvement due to the proposed multi-message HARQ is notable for high values of the throughput. In particular we observe that i) for any $\eta > 3$*, two transmissions of multi-message IR-HARQ performs better than the conventional IR-HARQ with infinite number of transmissions. Thus, if the throughput* $\eta > 3$ *is targeted, we may improve the performance and yet decrease the memory requirements at the receiver, and ii) increasing the number of transmission is also beneficial for multi-message IR-HARQ in the range of high SNR. For example, the SNR gap between* $\eta_\infty^{\text{mm}} = 3$ *and the ergodic capacity,* $\overline{C} = 3$*, is reduced by more than* 50% *when comparing to the gap between* $\eta_\infty = 3$ *and* $\overline{C} = 3$ *which is equal to* $5dB$*.*

V. Conclusions

In this work, aiming at the increase of the throughput of IR-HARQ transmission over block fading channel, we proposed and analyzed a new coding strategy tailored for HARQ protocol. In the proposed approach, the transmitter encodes an increasing number of messages in each round of HARQ. The rates of the messages are then optimized assuming existence of a multi-bits feedback channel which, on top of ACK/NACK signalling, conveys information about decoder state. The throughput of the resulting multi-message IR-HARQ is then compared to the conventional IR-HARQ.

The results indicate that significant gains can be obtained using the proposed coding strategy especially in the range of high throughput, where the conventional HARQ fails to offer any improvement with increasing number of transmission rounds. Namely, the proposed approach reduced the gap between the ergodic capacity and the conventional HARQ by more than 50%. The proposed encoding may be seen as a method to increase the throughput, or as a mean to diminish the memory requirements at the receiver; the price for the improvements is paid by a more complex joint encoding/decoding, whose practical aspects have yet to be analyzed in more details.

References

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

[2] P. Larsson, L. Rasmussen, and M. Skoglund, "Throughput analysis of hybrid-arq — a matrix exponential distribution approach," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 416–428, Jan. 2016.

[3] W. Lee, O. Simeone, J. Kang, S. Rangan, and P. Popovski, "HARQ buffer management: An information-theoretic view," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4539–4550, Nov. 2015.

[4] P. Larsson, L. K. Rasmussen, and M. Skoglund, "Throughput analysis of ARQ schemes in Gaussian block fading channels," *IEEE Trans. Commun.*, vol. 62, no. 7, pp. 2569–2588, Jul. 2014.

[5] J.-F. Cheng, Y.-P. Wang, and S. Parkvall, "Adaptive incremental redundancy," in *IEEE Veh. Tech. Conf. (VTC Fall)*, Orlando, Florida, USA, Oct. 2003, pp. 737–741.

[6] E. Visotsky, V. Tripathi, and M. Honig, "Optimum ARQ design: a dynamic programming approach," in *IEEE Inter. Symp. Inf. Theory (ISIT)*, Jun. 2003, p. 451.

[7] R. Liu, P. Spasojevic, and E. Soljanin, "On the role of puncturing in hybrid ARQ schemes," in *IEEE Inter. Symp. Inf. Theory (ISIT)*, Jun. 2003, p. 449.

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

[9] S. Pfletschinger and M. Navarro, "Adaptive HARQ for imperfect channel knowledge," in *2010 International ITG Conference on Source and Channel Coding (SCC)*, Jan. 2010, pp. 1 –6.

[10] S. M. Kim, W. Choi, T. W. Ban, and D. K. Sung, "Optimal rate adaptation for hybrid ARQ in time-correlated Rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 10, no. 3, pp. 968–979, Mar. 2011.

[11] 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.

[12] M. El Aoun, R. Le Bidan, X. Lagrange, and R. Pyndiah, "Multiple-packet versus single-packet incremental redundancy strategies for type-ii hybrid ARQ," in *6th International Symposium on Turbo Codes and Iterative Information Processing (ISTC), 2010*, 226-230, Ed., Sep. 2010.

[13] C. Hausl and A. Chindapol, "Hybrid ARQ with cross-packet channel coding," *IEEE Commun. Lett.*, vol. 11, no. 5, pp. 434–436, May 2007.

[14] D. Duyck, D. Capirone, C. Hausl, and M. Moeneclaey, "Design of diversity-achieving LDPC codes for H-ARQ with cross-packet channel coding," in *IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), 2010*, Sept. 2010, pp. 263–268.

[15] M. Jabi, A. El Hamss, L. Szczecinski, and P. Piantanida, "Multi-packet hybrid ARQ: Closing gap to the ergodic capacity," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 5191–5205, Dec. 2015.

[16] P. Popovski, "Delayed channel state information: Incremental redundancy with backtrack retransmission," in *IEEE Inter. Conf. Comm. (ICC)*, June 2014, pp. 2045–2051.
[17] K. Trillingsgaard and P. Popovski, "Block-fading channels with delayed CSIT at finite blocklength," in *IEEE Inter. Symp. Inf. Theory (ISIT)*, June 2014, pp. 2062–2066.
[18] M. Jabi, L. Szczecinski, M. Benjillali, and F. Labeau, "Outage minimization via power adaptation and allocation in truncated hybrid ARQ," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 711–723, Mar. 2015.
[19] L. Szczecinski and A. Alvarado, *Bit-Interlaved Coded Modulation : Fundamentals, Analysis and Design*. Wiley, 2015.
[20] M. Le Treust, L. Szczecinski, and F. Labeau, "Rate adaptation for secure HARQ protocols," in *IEEE Information Theory Workshop (ITW)*, Sep. 2013, pp. 1–5.
[21] D. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Athena Scientific, 2007, vol. 2.