# A Channel Representation Method for the Study of Hybrid Retransmission-Based Error Control

Leonardo Badia, *Member, IEEE,* Marco Levorato, *Student Member, IEEE,* and Michele Zorzi, *Fellow, IEEE*

***Abstract*— In this paper, we present a methodology to obtain a channel description tailored on performance evaluation for Incremental Redundancy Hybrid Automatic Repeat reQuest schemes. Such techniques counteract channel errors by using data coding and transmitting parts of the codeword over different channel realizations. We focus on coding performance models where the error probability is asymptotically zero if the channel parameters of these realizations fall within a given region. To map this region in a compact but still precise manner, we adopt a finite-state channel model. This approach is quite common in the literature; however, differently from existing work, we propose a novel method to derive efficient channel partitioning rules, i.e., a code-matched quantization of the channel state. Such a representation enables the use of accurate Markov models to study the system performance. Compared to existing channel representation methods, our proposed technique leads to a more accurate evaluation of higher layer statistics while at the same time keeping the computational complexity low.**

***Index Terms*— Channel coding, error correction, Markov processes, automatic repeat request, forward error correction, communication channels.**

## I. Introduction

THIS paper attempts to bridge fields which are of extreme interest for the scientific community working on wireless communication, yet are rarely investigated from a common viewpoint, namely, coding theory on the one hand, and error correction techniques applied to finite-state radio channel models on the other hand. We propose a channel representation methodology aimed at Hybrid Automatic Repeat reQuest (HARQ) [1], [2], based on recent advances in coding theory, where a renewed interest has arisen in deriving performance bounds for a broad class of practical codes [3]. In particular, it has been shown, for example, that for so-called *good* codes [4], such as Turbo codes [5] and Low–Density Parity–Check (LDPC) codes [6], the performance can be characterized by a threshold behavior. This leads to the definition of *reliable region* [7], i.e, the subspace of channel parameters where the error probability asymptotically vanishes as the codeword length becomes large.

Inspired by these findings, we propose in this paper a novel methodology for studying HARQ schemes, where we exploit this characterization of the code to determine a proper channel quantization. This allows to track the channel state through subsequent transmissions with an index spanning over a discrete set and therefore reduce the required system memory. Moreover, it enables an elegant and compact representation through a Finite-State Markov Chain (FSMC) of the HARQ process. We now describe some basics of the code performance bounds and HARQ, and point out the contribution of the present work.

Paper approved by L. K. Rasmussen, Editor for Iterative Detection Decoding and ARQ of the IEEE Communications Society. Manuscript received November 13, 2007, revised March 28, 2008. Part of this work has been presented at the Forty-Fifth Annual Allerton Conference on Communication, Control, and Computing.

L. Badia is with IMT Lucca Institute for Advanced Studies, piazza S. Ponziano 6, 55100 Lucca, Italy (e-mail: l.badia@imtlucca.it.).

M. Levorato and M. Zorzi are with the Dept. of Information Engineering, University of Padova, via Gradenigo 6/B, 35131 Padova, Italy.

Digital Object Identifier 10.1109/TCOMM.2008.xxxxx.

### A. Coding Performance Bound

Tight bounds on coding performance play a key role in the investigations about communication systems. New findings on this matter, leading to practical near-capacity-achieving codes, have drawn a considerable research effort aimed at deriving information theoretic coding performance bounds. Since the recent introduction of Turbo codes and the rediscovery of LDPC codes, the analytical derivation of tight bounds for good code ensembles has gained momentum. A code is said to be good when its block error probability asymptotically approaches zero as the codeword length increases if channel parameters fall within the reliable region [7]. As the structure of good codes is usually not available, bounds are often derived for ensembles of codes, i.e., classes of codes characterized by basic features of structure and construction, such as the distance spectrum and the input-output weight enumeration function [8], [9].

In [10]–[12], some tight bounds for good code ensembles with Maximum Likelihood (ML) decoding are proposed. An overview about these findings can also be found in [13]. Even though ML decoding requires unaffordable computational complexity, these bounds are useful to evaluate the ultimate performance limit. Moreover, performance evaluation through simulation shows that this limit may be approached by practical codes and decoding algorithms in several cases [10].

The success of iterative decoding algorithms, which harvest a considerable portion of the channel capacity with limited complexity, is one of the reasons of the success of Turbo–like codes. Recent work [14]–[17] computing the *noise threshold*, i.e., the reliable region boundary, for this decoding algorithm allows the evaluation of the performance loss with respect to optimal ML decoding.

### B. Background on Incremental Redundancy HARQ

The performance of plain ARQ can be improved by using channel coding so as to realize Hybrid ARQ schemes [1]. Especially, Incremental Redundancy (IR) implementations have

been widely employed for error control [2]. These techniques require codes with variable length, so that the first transmission utilizes the highest rate; then, additional redundancy bits are used whenever necessary for retransmissions. Several practical codes can work to this end; e.g., Turbo codes [5] fulfil the need for adaptive coding gain by means of puncturing schemes providing rate-compatible codes [18]–[20]. IR-HARQ schemes can be found in [21] and [22], where Turbo Codes and LDPC codes are used, respectively.

The growing interest in HARQ requires the derivation of bounds accounting for the fact that different parts of the codeword are sent over different channel realizations. In this case, the noise threshold becomes a multi-dimensional barrier that delimits the reliable region, and the asymmetry due to the diverse degrees of degradation suffered by the codeword symbols poses new difficulties. In [7], this is circumvented through the randomization of the symbol assignment to the codeword fragment to be sent over the different channels. Thus, the weight enumerator does not depend on a specific assignment, and the bound can be evaluated directly from the weight enumerator of the whole codeword. In [23], improved bounding techniques are proposed based on the random assignment as in [7].

The performance of Hybrid ARQ has been analyzed in several papers [24], [25] using a Markov chain to model the evolution of the transmission process. This approach may be complicated, since each state must convey all the information about the memory of the system, which may be very large [26]. Especially, tracing the channel evolution may require to store a huge amount of information. Thus, it is important to opt for a compact channel representation, able to preserve the tractability of the problem while obtaining meaningful results. A solution in this sense may be to apply a finite-state channel description, employing a proper channel quantization rule, i.e., representing the channel state with an index spanning over a discrete set. This methodology enables the creation of a FSMC exploiting the finite-state channel model. Markov techniques have gained foothold in both analytical and simulation frameworks, due to their good trade-off between accuracy and complexity [27], [28]. We remark that the most common approaches presented in the literature [29]–[31] to obtain a Markov representation of the channel use quantization criteria related to physical layer aspects, such as the equiprobability of the intervals.

### C. Contribution and Organization of the Paper

We propose to approach the issue presented above employing a novel methodology, i.e., a *Code-Matched* (CM) channel quantization, meaning that we explicitly take into account the coding performance when selecting how to partition the channel. We directly aim at giving an efficient approximation of the reliable region, so as to properly characterize the performance of IR-HARQ.

In particular, in this paper we give a detailed investigation for the specific case of an independent identically distributed (i.i.d.) channel where the HARQ system adopts a two-transmission limit, for which we are able to derive closed-form analytical results. Moreover, we instantiate our framework for an IR-HARQ scheme based on a Stop-and-Wait policy. The CM representation of the channel allows to efficiently evaluate the system performance by means of a proper Markov chain. Finally, we present numerical evaluations, where we bring examples of different codes, namely LDPC and Turbo codes, which can be used in the IR-HARQ scheme, and we quantify the goodness of our proposed approach in assessing IR-HARQ performance, compared to existing channel quantization techniques, such as equiprobable states [29]. These evaluations show that, compared to other techniques, the CM quantization obtains the same or a better characterization while employing a very limited number of states, thus achieving a channel description characterized by much lower complexity and/or memory requirements. Therefore, such a model can be extremely useful in both analytical investigations and simulation studies aimed at assessing the performance of IR-HARQ schemes.

The rest of this paper is organized as follows. In Section II we describe the IR-HARQ mechanism assumed in the analysis and we introduce some notation and terminology. Section III formalizes the proposed CM quantization model. Section IV focuses on the two-transmission case, deriving analytical results which allow for a compact characterization of the quantization thresholds and developing the FSMC to evaluate the HARQ performance. In Section V we present numerical evaluations which assess the superior match with the exact distribution of our channel representation with respect to uniform quantization, an interesting term of comparison since it is commonly used to obtain a discrete channel representation. Finally, we draw the conclusions in Section VI.

## II. System Model

This section outlines the assumptions used in the following to derive the analysis. It is subdivided into three parts, describing the generalities about IR-HARQ systems, the channel model, and the code description through the reliable region model.

### A. IR-HARQ Mechanism

An IR-HARQ system is characterized by the sequential transmission of *information frames*, each one of which is in turn associated with multiple *HARQ packets*. In practical cases, this is achieved by coding the information frame into a single long *codeword*, subdivided into multiple *fragments*, which are transmitted one at a time in a single HARQ packet. For this reason, in the following we will utilize the terms *frame* and *codeword* interchangeably, and analogously for *packet* and *fragment*. In order to keep the analysis simple, we postulate that all fragments are of the same size. This assumption can be removed with additional complications in the formulae, however the approach to follow is entirely similar. Also, we assume that even a single correctly received fragment is sufficient to decode the entire codeword. Sometimes, this situation is referred to as Type III ARQ [2].

When a packet arrives at the receiver's side, a feedback packet is sent back to the transmitter, indicating either positive (ACK) or negative acknowledgement (NACK). This feedback response refers to the whole information frame, since the

receiver can try to decode the codeword combining symbols contained in different fragments. Thus, an ACK means that the receiver was able to decode the frame based on all received HARQ packets associated with this frame (we will speak in this case of *frame resolution*), whereas a NACK means that the frame could not be decoded since the channel impairments exceeded the correction capability of the code formed by the set of currently received fragments. The key aspect of IR-HARQ is that a NACK does trigger a retransmission, but, unlike in other retransmission-based techniques, a physically different packet (though still associated with the same information frame) is transmitted. When this happens, we adopt a slight abuse of terminology by speaking of *frame retransmission*.

In this paper we focus on Stop-and-Wait (SW) HARQ, i.e., packets associated with the same information frame are sequentially transmitted, one at a time, after a feedback packet is received back at the transmitter. Extensions to other ARQ schemes, such as Go-Back-N and Selective Repeat, can be investigated within a conceptually similar framework. In SW ARQ, the transmission of packets associated with the same information frame goes on until either of these two conditions is met: (i) the set of received packets is sufficient to decode the frame; (ii) a maximum number $F$ of transmitted packets is reached without the receiver being able to decode the information frame, which is then discarded ($F$ may correspond to the total number of fragments generated for each codeword). In both cases, the transmission is then moved to another information frame. Thus, we will use the following notation. The *number of transmissions* experienced by a given frame will be noted with $\tau_T$, which can be between 1 and $F$. For any $k \in \{1, 2, \ldots, \tau_T\}$, we will refer to the codeword fragment received at the $k$th transmission as $w_k$. For completeness, we will use also $w_0$ to denote the codeword fragment (related to the previous frame) which has been transmitted previous to the first one of the current frame. Note that the case $\tau_T = F$ actually describes two different conditions, namely, the case where the frame is resolved exactly at the last available attempt, i.e., at the $F$th transmission, or the case where the frame is discarded. Even though the delay induced by the HARQ scheme on the transmission queue is the same, i.e., $F$ transmissions, the outcome is very different, since in the former case the frame is eventually resolved, in the latter it is not. To further distinguish between these two cases, we define another quantity called *resolution instant*, and denoted as $\tau_R$, as follows. If the frame is resolved at the $k$th transmission, the resolution instant is set to $k$. If the frame is *discarded*, $\tau_R$ is equal to $F+1$. Note that with this notation, $\tau_T = \min(\tau_R, F)$. In other words, $\tau_T$ and $\tau_R$ have the same value if the frame is resolved. Otherwise, if the frame is discarded, $\tau_T$ describes the delay experienced by the information stream, which is still equal to $F$ transmissions, whereas $\tau_R$ is conventionally set to $F+1$ to mark the difference with the case of a frame resolution at the last transmission attempt.

In the case $F = 1$, i.e., when a single transmission is allowed (a pure FEC situation), the analysis is straightforward. The outcome of the only packet transmission is either ACK or NACK according to the channel conditions and the correction capability of the code, which exhibits a binary (i.e., threshold-wise) behavior. If $F$ is increased to large values, an exact description of this process becomes cumbersome since it possibly includes the evaluation of $F$-dimensional thresholds.

### B. Channel Model

One important application of HARQ is to obtain reliable data transmission over a wireless fading channel. In the following we refer primarily to this scenario, where the transmission outcome depends on the Signal-to-Noise Ratio (SNR) at the receiver, even though the same rationale is directly applicable to other kinds of noisy channels. The channel model assumed in the rest of the paper is as follows. We assume a block flat fading channel, that is, each fragment experiences a single SNR coefficient, denoted with $s_k \in \mathbb{R}_+$ for the $k$th fragment, $k = 0, 1, \ldots, \tau_T$. This is similar to the block fading channel investigated for coded modulation in [32]; however, here the block length is a single frame, as in [33]. This representation can describe also HARQ applied to frequency-division schemes, for example transmission systems with frequency hopping or Orthogonal Frequency Division Multiplexing (OFDM) [22], [34].

We take a general description of the channel statistics, by considering the cumulative distribution function (cdf) and the probability density function (pdf) of the SNR values, which are denoted with $F_\Gamma$ and $f_\Gamma$, respectively. That is, if $\Gamma$ is the random variable describing the channel SNR,

$$F_\Gamma(s) = \text{Prob}\{\Gamma \leq s\}, \quad f_\Gamma(s) = \frac{\mathrm{d}F_\Gamma(s)}{\mathrm{d}s}. \tag{1}$$

Without loss of generality, we can assume that $f_\Gamma$ is strictly positive and therefore $F_\Gamma$ is increasing and invertible. If $F_\Gamma$ happens not to be invertible somewhere, i.e., there is an entire interval where it has the same value $x$, we can take any of these points as $F_\Gamma^{-1}(x)$.

However, differently from [32], our channel model is also able to account for correlated fading effects. To this end, we assume that the statistics of the SNR has the Markov property, which is a commonly adopted model to capture a correlated evolution of SNR values in consecutive frames [29], [31], [35]. Thus, we write $f_\Gamma(s_k|s_{k-1})$ to denote the conditional pdf of $s_k$ given that the SNR was equal to $s_{k-1}$ during the previous transmission. Under the assumption that the evolution of the SNR has the Markov property, the knowledge of $s_{k-1}$ allows to neglect any other condition on previous SNR values. In this way, our analysis also takes channel correlation into account. Clearly, if $f_\Gamma(s_k|s_{k-1}) = f_\Gamma(s_k)$ the channel model falls back to the case where the channel state is selected independently and at random from block to block, according to a known prior distribution, exactly as in [33].

### C. Code Description Through the Reliable Region

The reliable region model has been discussed in many papers appeared recently [10]–[17], [23]. In the following, we outline how this model is used in the present paper.

After $k$ transmissions of the same frame, the receiver bases the decoding of the codeword on all fragments received from $w_1$ up to $w_k$, whose SNR coefficients are $s_1, s_2, \ldots, s_k$. At the $k$th transmission, the reliable channel region, denoted with $\mathcal{R}(k)$, is a subset of $\mathbb{R}_+^k$ which contains the $k$-tuples of SNR

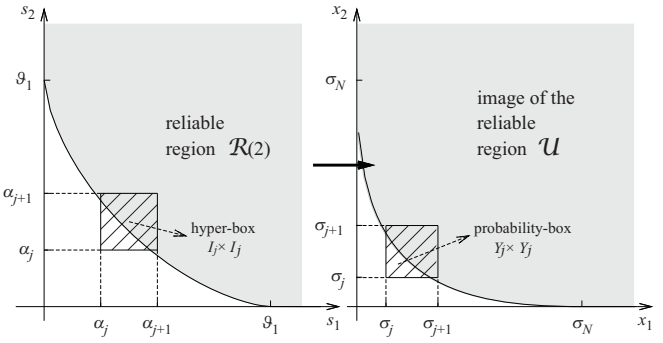



Fig. 1. A sample reliable region $\mathcal{R}(2)$ in $\mathbb{R}^2$, sample thresholds $\alpha_j$ and resulting interval and hyper-boxes on the left-hand side, and their corresponding images in the probability-domain, on the right-hand side.

coefficients where the failure probability becomes negligible if the packets sent are sufficiently large. Thus, the receiver is able to decode a codeword after the reception of its fragments $w_1, w_2, \ldots, w_k$ if $(s_1, s_2, \ldots, s_k) \in \mathcal{R}(k)$. A sample plot of the reliable region for $F = 2$, as well as other notations which can be useful for the reader to follow the discussions of the next sections, are reported in Fig. 1.

The exact specification of $\mathcal{R}(k)$ is determined by the code used, the decoding algorithm and the codeword fragments construction [7]. In any case, it is always verified that $(s_1, \ldots, s_{k-1}, s_k) \in \mathcal{R}(k)$ and $s'_k > s_k$ imply that $(s_1, \ldots, s_{k-1}, s'_k) \in \mathcal{R}(k)$. This fact immediately follows from the observation that if $s_k$ is a SNR value which is sufficient to decode the codeword at the $k$th transmission, given that the previous fragments have been received with SNR coefficients equal to $s_1, \ldots, s_{k-1}$, any SNR value greater than $s_k$ also enables frame resolution at the $k$th transmission, regardless of the particular shape of the reliable region.

Thanks to this property, we can use a representation of $\mathcal{R}(k)$ through a threshold function $\vartheta_k : \mathbb{R}_+^{k-1} \to \mathbb{R}_+$, defined as:

$$\vartheta_k(\mathbf{s}^{(k-1)}) = \inf\{s_k : (s_1, \ldots, s_{k-1}, s_k) \in \mathcal{R}(k)\}, \qquad (2)$$

where $\mathbf{s}^{(k-1)} = (s_1, s_2, \ldots, s_{k-1})$. In other words, the edge of $\mathcal{R}(k)$ is the curve identified by $\vartheta_k(\mathbf{s}^{(k-1)})$ in $\mathbb{R}_+^k$. Note that, when one transmission is considered, this curve degenerates to a single point $\vartheta_1 \in \mathbb{R}_+$, which is the value of $s_1$ associated with the (constant) SNR threshold to obtain correct decoding with a single fragment, and the reliable channel region $\mathcal{R}(1)$ corresponds to the interval $[\vartheta_1, +\infty[$.

Additionally, observe that $(s_1, \ldots, s_k) \in \mathcal{R}(k)$ also implies that $(s_1, \ldots, s_k, s_{k+1}) \in \mathcal{R}(k+1)$ for all $s_{k+1} \in \mathbb{R}_+$, since the fragments $w_1, w_2, \ldots, w_k$ were already sufficient to decode the codeword. This fact is referred to in the following as *success confirmation property*, and it describes that the transmission of a codeword can be dismissed after a success. Therefore, in the system under investigation, after the reception of a fragment $w_k$ the receiver is able to decode the packet if $s_k$ is above the threshold $\vartheta_k(\mathbf{s}^{(k-1)})$. In this case, no further fragment transmission is required (for this reason, the case where $\mathbf{s}^{(k)} \in \mathcal{R}(k)$ and $\mathbf{s}^{(k+1)} \in \mathcal{R}(k+1)$ never occurs in practice, but is considered here for completeness). Otherwise, another fragment $w_{k+1}$ is requested, which will be compared with threshold $\vartheta_{k+1}(\mathbf{s}^{(k)})$, and so on.

We denote with $\psi_{\tau_R, S|S_0}(j, s|s_0)$ the joint pdf of the resolution instant $\tau_R$ and SNR $S$ being equal to $j$ and $s$ respectively, when a frame is either resolved or discarded, conditioned on the SNR of the last transmission of the previous frame $S_0$ being equal to $s_0$. Similarly, we denote with $\psi_S(s)$ the pdf of the SNR $S$ being equal to $s$ in the global case of frame resolution or discarding. If $\mathcal{X} \subseteq \mathbb{R}_+^k$, we define $\mathbb{1}(\mathbf{s}^{(k)}, \mathcal{X})$ as equal to 1 if $\mathbf{s}^{(k)} \in \mathcal{X}$, and 0 otherwise. We have:

$$\psi_{\tau_R, S|S_0}(j, s|s_0) =$$

$$\begin{cases} f_\Gamma(s|s_0)\, \mathbb{1}\big((s), \mathcal{R}(1)\big) & \text{if } j = 1 \\ \displaystyle\int_{\mathbb{R}_+^{j-1} \setminus \mathcal{R}(j-1)} f_\Gamma(s|s_{j-1})\, \mathbb{1}\big((\mathbf{s}^{(j-1)}, s), \mathcal{R}(j)\big) \\ \quad \displaystyle\prod_{\ell=1}^{j-1} f_\Gamma(s_\ell|s_{\ell-1})\, \mathrm{d}s_\ell & \text{if } 2 \le j \le F \\ \displaystyle\int_{\mathbb{R}_+^{F-1} \setminus \mathcal{R}(F-1)} f_\Gamma(s|s_{F-1})\, \mathbb{1}\big((\mathbf{s}^{(F-1)}, s), \mathbb{R}_+^F \setminus \mathcal{R}(F)\big) \\ \quad \displaystyle\prod_{\ell=1}^{F-1} f_\Gamma(s_\ell|s_{\ell-1})\, \mathrm{d}s_\ell & \text{if } j = F+1 \end{cases}$$

$$\Rightarrow \psi_S(s) = \int_0^{+\infty} \psi_S(s_0) \sum_{j=1}^{F+1} \psi_{J, S|S_0}(j, s|s_0)\, \mathrm{d}s_0\,. \qquad (3)$$

Taking $\beta(s, s_0) = \displaystyle\sum_{j=1}^{F+1} \psi_{J, S|S_0}(j, s|s_0)$, we can write (3) as

$$\psi_S(s) = \int_0^{+\infty} \psi_S(s_0) \beta(s, s_0) \mathrm{d}s_0\,, \qquad (4)$$

from which it is clear that $\psi_S(s)$ is an eigenfunction of the transform operation, which maps a generic function $\zeta : \mathbb{R} \to \mathbb{R}$ into its transform $\widehat{\zeta}$, given by

$$\widehat{\zeta}(s) = \int_0^{+\infty} \zeta(s_0) \beta(s, s_0) \mathrm{d}s_0\,. \qquad (5)$$

This enables a recursive strategy to determine $\psi_S(s)$ as the limit of a sequence $\{\zeta_n(s)\}_{n \in \mathbb{N}}$ of functions where $\zeta_0(s) = f_\Gamma(s)$ and $\zeta_{\ell+1}(s) = \widehat{\zeta_\ell}(s)$, as per (5).

In this manner, the probability that a $k$-tuple of SNR values *associated with the same frame* belong to a given region $\mathcal{X} \subseteq \mathbb{R}_+^k$ is determined as:

$$\int_{\mathbb{R}_+} \psi_S(s_0) \int_{\mathcal{X}} f_\Gamma(s_1|s_0) f_\Gamma(s_2|s_1) \ldots f_\Gamma(s_k|s_{k-1})\, \mathrm{d}s_k \ldots \mathrm{d}s_2\, \mathrm{d}s_1\, \mathrm{d}s_0\,. \qquad (6)$$

## III. Code-Matched Channel Quantization and Construction of the Markov Chain

To exactly evaluate the process described above, a very large amount of information is required at each step. In fact, to determine whether the frame could be acknowledged after the transmission of fragment $w_{k+1}$, the $k$–dimensional vector $\mathbf{s}^{(k)}$ must be traced. However, high complexity and memory requirements are implied to accurately track the evolution of a vector of continuous variables. Henceforth, it is meaningful

to consider a quantization of the SNR to enable a finite-state channel representation, where each state represents an interval of SNR values, and which can be used in an FSMC context.

Thus, we partition $\mathbb{R}_+$ into $N+1$ non overlapping adjacent intervals $I_0, I_1, ..., I_N$. These intervals are identified by choosing $N$ proper threshold values $\alpha_1, \alpha_2, \ldots, \alpha_N$, so that $\alpha_j < \alpha_{j+1}$ and $I_j = [\alpha_j, \alpha_{j+1}[$. For consistency, we take $\alpha_0 = 0$, $\alpha_{N+1} = +\infty$. The purpose of this partition is to represent the SNR with a finite number of states, in order to use a discrete description of the channel. We will talk in the following of a *quantized channel*, where we no longer know the exact SNR values, but only which region the SNR falls within. In fact, according to this representation, any sequence of SNR values $\mathbf{s}^{(k)}$ is described with a sequence of discrete values in $\mathbb{Z}_{N+1}^k$, where $\mathbb{Z}_{N+1} = \{0, 1, \ldots, N\}$. Let $d(\cdot)$ be the corresponding function mapping $\mathbb{R}_+$ to $\mathbb{Z}_{N+1}$ such that $d(s_k) = \arg_j(s_k \in I_j)$. With a slight abuse of notation, we write $\mathbf{d}^{(k)} = d(\mathbf{s}^{(k)}) = (d(s_1), d(s_2), \ldots, d(s_k))$. Thus, the $k$-tuple $\mathbf{d}^{(k)}$ indicates that for the $j$th received SNR, $s_j \in I_{d_j}$, for all $j = 1, 2, \ldots, k$. Moreover, we define the *hyper-box* $\mathcal{I}(\mathbf{d}^{(k)}) = I_{d(s_1)} \times I_{d(s_2)} \times \cdots \times I_{d(s_k)} \subseteq \mathbb{R}_+^k$.

The channel quantization can be used to determine an FSMC describing the channel as follows. The state space is $\mathbb{Z}_{N+1}$, and each state $i$ can be shown [30] to have a steady-state probability $\pi_i$ equal to

$$\pi_i = \int_{\alpha_i}^{\alpha_{i+1}} f_\Gamma(s) \mathrm{d}s \,. \tag{7}$$

In a similar manner, the transition probability from state $i$ to state $j$, denoted with $p_{ij}$, can be obtained as

$$p_{ij} = \frac{\int_{\alpha_i}^{\alpha_{i+1}} f_\Gamma(s) \int_{\alpha_j}^{\alpha_{j+1}} f_\Gamma(t|s) \, \mathrm{d}t \, \mathrm{d}s}{\int_{\alpha_i}^{\alpha_{i+1}} f_\Gamma(s) \mathrm{d}s} \,. \tag{8}$$

The finite-state model outlined above can be used to evaluate the performance of HARQ. For graphical aid, refer to Fig. 1. If it is known that a sequence of $k$ HARQ packets has been experienced SNR conditions which, according to the discrete representation of the channel, are represented by vector $\mathbf{d}^{(k)}$, it follows that the overall SNR vector $\mathbf{s}^{(k)}$ falls within the hyper-box $\mathcal{I}(\mathbf{d}^{(k)})$. By comparing the relative placement of $\mathcal{I}(\mathbf{d}^{(k)})$ and the reliable region $\mathcal{R}(k)$, it is possible to infer the outcome of the reception of the fragments $w_1, w_2, \ldots, w_k$, i.e., whether the frame is acknowledged or not. If $\mathcal{I}(\mathbf{d}^{(k)})$ is entirely within or outside $\mathcal{R}(k)$, the frame is correspondingly acknowledged or not, respectively. However, when $\mathcal{I}(\mathbf{d}^{(k)})$ only partially overlaps with $\mathcal{R}(k)$, the hyper-box will be assumed to correspond to frame resolution or not according to a ML criterion. This means to check whether

$$\int_{\mathbb{R}_+} \psi_S(s_0) \int_{\mathcal{I}(\mathbf{d}^{(k)}) \cap \mathcal{R}(k)} f_\Gamma(s_1|s_0) f_\Gamma(s_2|s_1) \ldots \\ \ldots f_\Gamma(s_k|s_{k-1}) \, \mathrm{d}s_k \ldots \mathrm{d}s_2 \, \mathrm{d}s_1 \, \mathrm{d}s_0 \,. \tag{9}$$

is larger than

$$\int_{\mathbb{R}_+} \psi_S(s_0) \int_{\mathcal{I}(\mathbf{d}^{(k)}) \setminus \mathcal{R}(k)} f_\Gamma(s_1|s_0) f_\Gamma(s_2|s_1) \ldots \\ \ldots f_\Gamma(s_k|s_{k-1}) \, \mathrm{d}s_k \ldots \mathrm{d}s_2 \, \mathrm{d}s_1 \, \mathrm{d}s_0 \,. \tag{10}$$

If this is true, we say that $\mathcal{I}(\mathbf{d}^{(k)})$ is a *resolved* hyper-box, otherwise, we call it an *unresolved* hyper-box. The reliable region is therefore approximated, in the quantized representation, by the union set $\mathcal{J}(k)$ of the resolved hyper-boxes, i.e.,

$$\mathcal{J}(k) = \bigcup_{\substack{\text{resolved} \\ \mathcal{I}(\mathbf{d}^{(k)})}} \mathcal{I}(\mathbf{d}^{(k)}). \tag{11}$$

Thus, we have an overall approximate description of the whole HARQ process where the frame is either acknowledged or not, after $k$ of its packets have been received, according to $\mathcal{I}(\mathbf{d}^{(k)})$ being resolved or not. This representation introduces an approximation every time both (9) and (10) are non-zero. However, observe that the hyper-boxes maintain certain properties of the reliable region. For example, due to the monotonicity of $f_\Gamma$ and $\vartheta_k$ the success confirmation property holds also for $\mathcal{J}(k)$, i.e., if $\mathcal{I}(\mathbf{d}^{(k-1)})$ is resolved, $\mathcal{I}(\mathbf{d}^{(k)})$ is resolved as well. Our objective is to reduce the decision errors introduced by this representation, e.g., that an SNR $k$-tuple $\mathbf{s}^{(k)}$ would imply retransmission, but it corresponds to a resolved hyper-box $\mathcal{I}(\mathbf{d}^{(k)})$, or conversely that the parameters $\mathbf{s}^{(k)}$ would allow for frame resolution but the ML criterion for $\mathcal{I}(\mathbf{d}^{(k)})$ gives retransmission.

Any given choice of the set of thresholds $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ implies quantization errors in the description of the SW HARQ process using the quantized channel. For example, the resolution instant of a given frame can be erroneously identified as $k$, whereas the frame may be, in fact, resolved at a previous transmission, at a later one, or even discarded (i.e., the resolution instant is $\tau_R = F + 1$).

In general, we are interested in avoiding these errors. A *code-matched* (CM) quantization is a choice of the thresholds which is made with this purpose in mind. Actually, there are several possibilities to define a CM quantization. None of them is, strictly speaking, optimal in an absolute sense; their optimality depends on which definition of the quantization error is chosen. In the following, we will give two possible definitions (i.e., two different error terms) and we will elaborate a CM quantization which tries to address both of them. However, we remark that some variations in this sense are possible, leading to minor differences. The important concept is that the choice of the threshold must be made with the aim of achieving an efficient channel description where quantization errors are small.

First, we define the *quantization error probability on the resolution instant*, denoted with $Q_R$, as the probability that the resolution instant of a frame differs in the true and the quantized channel. This is a quantity that one may want to minimize, as it describes the occurrence of a wrong evaluation in the frame outcome. If a given frame is transmitted for the $k$th time, and in all the $k-1$ previous transmissions both the true and the quantized channels have identified all the sets $\mathbf{s}^{(j)} = (s_1, s_2, \ldots, s_j)$, with $j \leq k-1$, as implying retransmission, we denote with $A_k$ the probability that both the true and the quantized channels identify now the frame as resolved, and with $B_k$ the probability that both of them identify it as implying another retransmission. To formalize this, define $\mathcal{Q}(k) = \mathbb{R}_+^k \setminus (\mathcal{R}(k) \cup \mathcal{J}(k))$, i.e., $\mathcal{Q}(k)$ is the intersection of the complementary sets of the reliable region

$\mathcal{R}(k)$ and the estimated reliable region $\mathcal{J}(k)$. Note that, due to the success confirmation property, $\mathcal{Q}(k)$ can be shown to be a subset of $\mathcal{Q}(k-1)\times\mathbb{R}_+$. Thus,

$$\begin{aligned} A_k &= \int_{\mathbb{R}_+} \psi_S(s_0) \int_{\left(\mathcal{R}(k)\cap\mathcal{J}(k)\right)\cap\left(\mathcal{Q}(k-1)\times\mathbb{R}_+\right)} f_\Gamma(s_1|s_0) \\ &\quad f_\Gamma(s_2|s_1)\ldots f_\Gamma(s_k|s_{k-1})\,\mathrm{d}s_k\ldots\mathrm{d}s_2\,\mathrm{d}s_1\,\mathrm{d}s_0 \\ B_k &= \int_{\mathbb{R}_+} \psi_S(s_0) \int_{\mathcal{Q}(k)\cap\left(\mathcal{Q}(k-1)\times\mathbb{R}_+\right)} f_\Gamma(s_1|s_0) \\ &\quad f_\Gamma(s_2|s_1)\ldots f_\Gamma(s_k|s_{k-1})\,\mathrm{d}s_k\ldots\mathrm{d}s_2\,\mathrm{d}s_1\,\mathrm{d}s_0 \\ &= \int_{\mathbb{R}_+} \psi_S(s_0) \int_{\mathcal{Q}(k)} f_\Gamma(s_1|s_0) f_\Gamma(s_2|s_1)\ldots \\ &\quad \ldots f_\Gamma(s_k|s_{k-1})\,\mathrm{d}s_k\ldots\mathrm{d}s_2\,\mathrm{d}s_1\,\mathrm{d}s_0\,. \end{aligned} \tag{12}$$

Therefore, we have

$$Q_R = \sum_{k=1}^{F}(1-A_k-B_k)\prod_{j=1}^{k-1}B_j\,. \tag{13}$$

This equation evaluates $Q_R$ by summing all cases where the true and the quantized channel agree for $k-1$ transmissions, and disagree on the $k$th, where $k$ is summed from 1 to $F$. Note that both channels agree that transmission $k$ takes place only with probability $B_1B_2\ldots B_{k-1}$, which is why we multiply by the last product term. In other words, provided that both representations agree up to $k-1$ transmissions, we have three possibilities for the $k$th transmission: (i) they both agree that the frame is resolved, which occurs with probability $A_k$; (ii) they disagree, which happens with probability $(1-A_k-B_k)$; (iii) they agree that the frame requires another transmission, which occurs with probability $B_k$. In the first two cases, the agreement, for case (i) or the disagreement, for case (ii), is reflected on the final outcome. In the last case, the fact that the true and the quantized channel agree on the $k$th transmission does not imply that they agree also on the final outcome of the frame, unless $k=F$ (in which case they agree the frame is discarded). If $k<F$ we need another transmission to see whether the two representations agree, which can be seen by increasing $k$ by 1 and iterating the reasoning above.

Note that in evaluating $Q_R$, both false positive and false negative (the true channel is in error but the quantized channel calls the frame as correct or vice versa) are accounted for with the same weight. However, it is straightforward to modify (13) to consider different weights.

Similarly to $Q_R$ we can define another error term, namely the *quantization error probability on the number of transmissions*, denoted with $Q_T$, as the probability that the number of transmissions experienced by a frame, $\tau_T$, differs in the true and the quantized channel. To evaluate $Q_T$, we modify (13) by taking the upper limit of the summation as $F-1$, i.e.,

$$Q_T = \sum_{k=1}^{F-1}(1-A_k-B_k)\prod_{j=1}^{k-1}B_j\,. \tag{14}$$

Recalling the definitions of $\tau_R$ and $\tau_T$, we see that $Q_R$ (resp., $Q_T$) counts the probability that the evaluation of $\tau_R$ (resp., $\tau_T$) made through the quantized channel is erroneous. In other words, $Q_R$ indicates whether the outcome of the frame is correctly evaluated or not, whereas $Q_T$ denotes whether the delay evaluation is right. In the latter case, quantization errors on the last transmission are neglected as they do not change the delay experienced by the frame (though they change its outcome). In fact, a discarded frame or a frame which is resolved at the last possible transmission both experience a delay equal to $F$. Thus, the minimization of $Q_T$ is appropriate to assess the delay performance of the HARQ process. To evaluate the throughput, instead, one should seek to minimize $Q_R$. Also, intermediate choices (e.g., minimizing a combination of $Q_T$ and $Q_R$) are possible as well. In this sense, we tried many possibilities, and the ones described next seems to be a reasonable choice for the cases under exam in this paper. Actually, due to the similarity of the definitions of these metrics, other approaches are possible with almost negligible differences.

In the following, we will adopt a hybrid criterion. We define a CM quantization adopting this rule: minimize $Q_T$ first, by using some of the thresholds, and then minimize $Q_R$ by using the remaining thresholds. This is motivated by the practical cases considered in the following sections, where we consider systems which can have $Q_T=0$ if some of the thresholds (actually, one) are properly chosen. Thus, the CM condition translates into minimizing $Q_R$ given that $Q_T=0$.

Finally, note that another option, which is still compatible with our framework, is to weigh more the errors when the difference between the resolution instant and its estimate is higher. However, since in most of the present paper we focus on the case $F=2$, using different weights would only slightly affect what is discussed in the following. Anyway, it is another point to account for, especially if the correct evaluation of the delay is the most important aspect of the HARQ analysis.

## IV. Application to $F=2$, i.i.d. Channel

As a practical case for the evaluation of our proposed CM quantization technique, we consider a simple SW HARQ system in which $F=2$ and different channel realizations are characterized by i.i.d. SNR processes. This means that $f_\Gamma(s_k|s_{k-1})=f_\Gamma(s_k)$ for all $k$.

We assume that the reliable region $\mathcal{R}(2)$ is convex and symmetric with respect to permutations of the coordinates, i.e., $(s_1,s_2)\in\mathcal{R}(2)$ implies that $(s_2,s_1)\in\mathcal{R}(2)$ as well. Also, we assume that $\vartheta_2$, describing the edge of $\mathcal{R}(2)$, is a strictly decreasing function. These hypotheses are well verified for the specific choices used later in the results' section, which represent realistic scenarios, and are reasonable for most practical cases [7], [23], [36].

### A. Preliminary Remarks

For any $j$ in $\mathbb{Z}_{N+1}$, consider the $(N+2)$-tuple of labels (equal to either "resolved" or "unresolved") obtained by taking the outcomes of the ML criterion reported in (9) and (10), applied to the following hyper-boxes (and, correspondingly, to their associated probability-boxes): $I_j$ alone, compared to the reliable region $\mathcal{R}(1)$, as the first label; $I_j\times I_0$, $I_j\times I_1$, …, $I_j\times I_N$, compared to $\mathcal{R}(2)$ as the other $N+1$ labels.

In the following, we will refer to this set of labels as the *character* of the $j$th state of the quantized channel. Note that in the definition the products can be reverted, thus going row-by-row instead of column-by-column. Finally, we say that two states are adjacent if their indices differ by 1.

Because of the success confirmation property, if the character of state $j$ contains a "resolved" for $I_j$ alone, the same is true for all other labels. Instead, if $I_j$ alone is unresolved, we have that, since the curve $\vartheta_2$ is decreasing, there exists a value $\chi(j) \in \mathbb{Z}_{N+1}$ such that all $I_j \times I_k$ are resolved (resp. unresolved) if $k \geq \chi(j)$ (resp. $k < \chi(j)$). In the following, $\chi(j)$ will be called the *height* of the $j$th state. Evidently, the height $\chi(j)$ is a non-increasing (integer) function of $j$. Since we assume that the reliable region is symmetric with respect to permutations of the coordinates, it holds that the height of the $\chi(j)$th state is $j$.

To find a suitable quantization, we observe that partitioning region $\mathcal{R}(1)$ with a single threshold point at $\vartheta_1$ is a choice that makes the quantized model coincide with the true channel. For the two transmission case, an exact representation of $\mathcal{R}(1)$ implies that $Q_T = 0$. Thus, the code match criterion discussed in the previous section, i.e., to impose $Q_T = 0$ and then minimize $Q_R$, can be achieved if we choose one threshold as $\vartheta_1$ and then we use the other $N-1$ thresholds to further minimize $Q_R$. If we apply (13) to derive the CM condition, this corresponds to choosing the thresholds so as to minimize $A_2 + B_2$, as defined by (12). That is, if $\mathcal{V} = (\mathcal{R}(2) \setminus \mathcal{J}(2)) \cup (\mathcal{J}(2) \setminus \mathcal{R}(2))$,

$$\min \int_{\mathcal{V}} f_\Gamma(s_1) f_\Gamma(s_2)\, \mathrm{d}s_2\, \mathrm{d}s_1\,. \tag{15}$$

At this point, to simplify the notation, it is useful to employ function $F_\Gamma(s) = \text{Prob}\{\gamma < s\}$, mapping $\mathbb{R}_+$ into $[0,1]$, for a change of variables $x_1 = F_\Gamma(s_1)$, $x_2 = F_\Gamma(s_2)$. This operation is depicted in Fig. 1. The reliable region $\mathcal{R}(2)$ can be described by its image $\mathcal{U} = \Phi(\mathcal{R}(2))$, where $\Phi(s_1, s_2) = (F_\Gamma(s_1), F_\Gamma(s_2))$. The region $\mathcal{U}$ can in turn be represented through a function $\Theta : [0,1] \to [0,1]$, so that $x_2 = \Theta(x_1) = F_\Gamma\Big(\vartheta_2\big(F_\Gamma^{-1}(x_1)\big)\Big)$. For the sake of analytical tractability, in the following we assume that $\Theta$ is convex and symmetric, as $\vartheta_2$ was. [1]

Condition (15) can then be re-arranged into

$$\min \int_{\mathcal{W}} \mathrm{d}F_\Gamma(s_2)\, \mathrm{d}F_\Gamma(s_1) = \min \int_{\mathcal{W}} \mathrm{d}x_2\, \mathrm{d}x_1\,, \tag{16}$$

where $\mathcal{W} = \Phi(\mathcal{V})$. We refer to $\mathcal{W}$ as the *error region*, since it describes the probability of those cases contributing to $Q_R$. Now, we simply need to minimize the area of the error region $\mathcal{W}$, which depends on the shape of function $\Theta$ and on the placement of the thresholds.

Similarly to the mapping of the reliable region, in the following we will consider, instead of the SNR thresholds $\alpha_k$, their outputs through $F_\Gamma$, i.e., $\sigma_k = F_\Gamma(\alpha_k)$. The values of $\sigma_k \in [0,1]$ will be referred to as *probability thresholds* of the quantization. Due to the monotonicity of $F_\Gamma$, the probability thresholds $\sigma_k$ and the SNR thresholds $\alpha_k$ give an equivalent description of the quantization process.

We denote with $Y_j$ the image through $F_\Gamma$ of the $j$th interval $I_j$, so that the probability thresholds partition $[0,1]$ into $Y_0, Y_1, \ldots, Y_N$. For $j, \ell \in \mathbb{Z}_{N+1}$, the hyper-box $I_j \times I_\ell$ of the two-dimensional space of SNRs is similarly mapped by $\Phi$ into a *probability-box* $Y_j \times Y_\ell$ in $[0,1]^2$. The assignment of labels such as "resolved" or "unresolved" for a hyper-box $I_j \times I_\ell$ identically holds for the corresponding probability-box $Y_j \times Y_\ell$, and so do the definitions of "character" and "height."

### B. Derivation of the CM Quantization

Interestingly, for the case under study, the optimal thresholds obey a general closed-form expression. In this subsection, we demonstrate some useful preliminary result and finally we prove this statement.

*Lemma 1:* Given $N$ thresholds, it is always possible to keep them fixed and add another threshold to obtain a quantization with $N+1$ thresholds where $Q_R$ is strictly smaller.

The intuition of the proof is quite straightforward. A formal proof is reported in Appendix I.

*Lemma 2:* If $Q_R$ is minimized, all the states must have a different character.

*Proof:* Due to the decreasing behavior of the curve $\Theta$ and the success confirmation property, if two non-adjacent states have the same character, all the intermediate states have the same character as well. Thus, it is sufficient to prove that no adjacent states can have the same character. This is intuitively immediate, as if we assume to have two "redundant" adjacent states $j$ and $j+1$ with the same character, we can merge them so as to obtain an $N-1$ threshold placement where the quantization error probability is kept unchanged. Then, we exploit Lemma 1 to obtain a quantization with $N$ thresholds and strictly smaller quantization error probability. Thus, the initial threshold choice could not be minimal in $Q_R$. ■

As a corollary to Lemma 2, at most one threshold may fall in $[\vartheta_1, +\infty[$, and at most one state can be characterized by success after the first transmission. Thus, if one of the thresholds equals $\vartheta_1$ it must be the largest threshold $\alpha_N$. Moreover, observe that Lemma 2 holds true only in the i.i.d. case, whereas in the correlated case using "redundant" states, i.e., with the same character, may be desirable, since the channel quantization represents not only the outcome of the HARQ process, but also the channel transitions, which are possibly influenced by the underlying correlation. In this case, two states may have the same outcomes, though having different transition probabilities to the next state.

*Theorem 3:* Take two adjacent states in the quantization minimizing $Q_R$. Either they have a different character for the first transmission, or their heights differ exactly by 1.

*Proof:* If two adjacent states, say $j$ and $j+1$, differ in the labels of $I_j$ and $I_{j+1}$, there is nothing to prove. Otherwise, due to Lemma 2, they both identify the first transmission as unresolved, and $\chi(j+1)$ must be *strictly* lower than $\chi(j)$. Now, we proceed by contradiction. Were $\chi(j) - \chi(j+1) \geq 2$, states $\chi(j+1)$ and $\chi(j)$ would not be adjacent, so there would

[1] It may happen that region $\mathcal{U}$ is no longer convex if the channel conditions are bad (low average SNR) and correlation is very strong. However, this is just a border effect (the region $\mathcal{U}$ "bends" only around 1) and does not affect the following analysis apart from introducing more cumbersome equations. Moreover, we observe that for most cases of interest, e.g., for Rayleigh fading with a sufficiently high average SNR (0 dB is sufficient), the curve $\Theta$ is indeed convex.

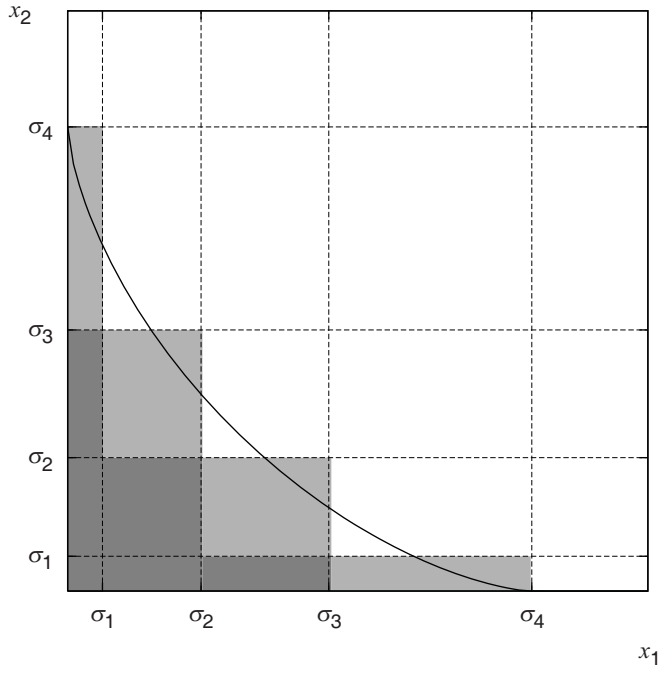



Fig. 2. Visual comparison of $i$-approach (reliable region is the white area) and $x$-approach (reliable region is the union of white and light grey areas).

be an intermediate state $\ell$, with $\chi(j+1) < \ell < \chi(j)$; e.g., take $\ell = \chi(j+1)+1$. As $\chi(\chi(j)) = j$, the height $\chi(\ell)$ would be an integer strictly smaller than $j+1$, but also strictly larger than $j$, which is a contradiction. ■

Another consequence of Theorem 3 is that only two strategies are possible to determine the characters of the states. In the first one, which in the following will be referred to as *internal approach*, or i-approach for short, the height of the $j$th column is $\chi(j) = N - j$. In the second one, referred to as *external approach*, or x-approach, the height of the $j$th column is $\chi(j) = N - j - 1$. In Fig. 2 we plot a graphical comparison of these two approaches, to show their difference. The i-approach corresponds to approximating the region $\mathcal{U}$ with the white area only, whereas the x-approach considers both white and grey boxes as part of the reliable region. As a side remark, observe that the x-approach does not violate Lemma 2, since the states $N-1$ and $N$ have a different label (unresolved and resolved, respectively) for the first transmission.

We now derive the thresholds which locally minimize $Q_R$ (i.e., the area of the error region $\mathcal{W}$) for both approaches. Further, we show that the global minimum is achieved by the i-approach.

*Theorem 4:* For the i-approach, $Q_R$ is minimized if

$$\Theta(\sigma_i) = \frac{\sigma_{N-i} + \sigma_{N-i+1}}{2}, \quad \text{for } i = 1, \ldots, N-1\,. \quad (17)$$

The proof is reported in Appendix II.

*Theorem 5:* For the x-approach, $Q_R$ is minimized if

$$\Theta(\sigma_i) = \frac{\sigma_{N-i-1} + \sigma_{N-i}}{2}, \quad \text{for } i = 1, \ldots, N-1\,. \quad (18)$$

We do not explicitly demonstrate this statement, since it can be proven similarly to Theorem 4. Not only can the proof of Theorem 5 be obtained along the lines of Appendix II, but also the next Theorem justifies that a detailed proof is irrelevant, as the x-approach is sub-optimal.

*Theorem 6:* The quantization error probability on the resolution instant resulting from choosing the optimal thresholds of the i-approach is lower than the value achieved by the optimal thresholds of the x-approach.

The proof is reported in Appendix III.

Due to Theorem 6, it is necessary to consider (17) only. Using this expression, it is easy to find the optimal SNR thresholds $\alpha_1, \alpha_2, \ldots, \alpha_N$ via simple numerical methods. Moreover, (17) can be seen as a system of $N-1$ equations with $N-1$ unknowns, i.e., $\sigma_1, \sigma_2, \ldots, \sigma_{N-1}$, as $\sigma_N$ is already chosen as $F_\Gamma(\vartheta_1)$, and due to the monotonicity property of $\Theta$ and $F_\Gamma$ the uniqueness of the solution is guaranteed. Thanks to this theoretical finding, it is possible to identify an efficient partitioning method of the SNR values which is matched to the characteristics of the code and is therefore more suitable to describe the HARQ process through a Markov model.

### C. Resulting SW HARQ Scheme

If $F = 2$ the resolution instant of a new SW HARQ frame can equal 1, 2, or 3, meaning that the frame is resolved after the first transmission, resolved after the second transmission, or discarded, respectively. Formally, the resolution instant $\tau_R$ equals $j \in \{1, 2, 3\}$ with probability $\eta(j)$, equal to

$$\begin{aligned}
\eta(1) &= \text{Prob}\{s \geq \vartheta_1\} = 1 - F_\Gamma(\vartheta_1) \\
\eta(2) &= \text{Prob}\{s_1 < \vartheta_1, s_2 \geq \vartheta_2(s_1)\} \\
&= \int_0^{\vartheta_1} \Big(1 - F_\Gamma\big(\vartheta_2(s_1)\big)\Big) f_\Gamma(s_1)\mathrm{d}s_1 = \int_{\mathcal{U}'} \mathrm{d}x_1 \mathrm{d}x_2
\end{aligned}$$

where $\mathcal{U}' = \mathcal{U} \cap ([0, \sigma_N] \times [0, 1])$, and $\eta(3) = 1 - \eta(1) - \eta(2)$.

An approximated evaluation of these quantities can be achieved by using a quantized channel instead of the true channel. Note that we previously defined a *channel chain*, i.e., an FSMC describing the evolution of the quantized channel. This chain infers a different FSMC, called the *HARQ chain*, representing the HARQ process. In particular, if the quantized channel is obtained with $N$ thresholds, i.e., the channel chain has $N+1$ states, we can derive a HARQ chain with $N+2$ states. In the HARQ chain, there are $N$ states $0, 1, 2, \ldots, N-1 \in \mathbb{Z}_N$ to track the channel condition after a retransmission, plus two additional states, named $\omega$ and $\delta$, corresponding to frame resolution or discard, respectively. Observe that there is no state $N$ in the HARQ Markov chain, since if the channel chain is in state $N$, the frame is necessarily resolved so the HARQ Markov chain is in state $\omega$.

States $\omega$ and $\delta$ correspond to the first transmission of a frame. Since the channel is i.i.d., and both resolution and discarding enable the transmission of a new frame, the exits from these states are identical. In particular, if the channel is in state $N$, the new frame is immediately resolved, thus we enter state $\omega$. Otherwise, we enter a state in $\mathbb{Z}_N$, according to the channel state. The states in $\mathbb{Z}_N$ describe the SNR of an erroneous first transmission, after which the frame is necessarily resolved or discarded, so that states $\omega$ or $\delta$ are entered, correspondingly. This representation gives an exact evaluation of $\eta(1)$, whereas $\eta(2)$ is estimated only through the chain state, which is a discrete value, whereas a continuous value, i.e., the SNR $s_1$ of the first packet transmitted, would

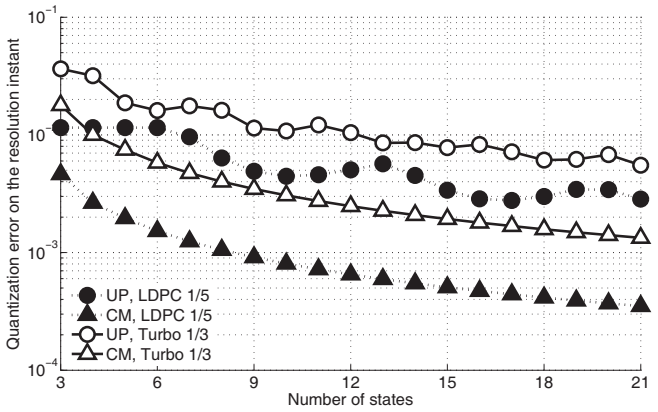



Fig. 3. Quantization error probability on the resolution instant, $Q_R$, of the CM and UP quantization methods versus the number of channel states, $R = 2.6$ bps/Hz, $\gamma_0 = 6$ dB, two transmissions.

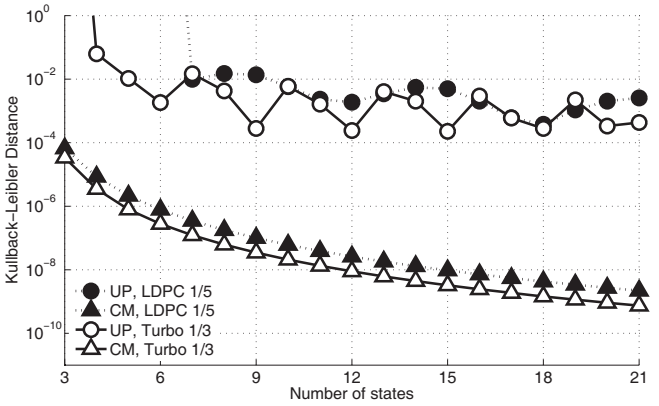



Fig. 4. Kullback Leibler divergence of the distribution of $\tau_R$ for the CM and UP quantization methods versus the number of channel states, $R = 2.6$ bps/Hz, $\gamma_0 = 6$ dB, two transmissions.

be required. On the other hand, using a discrete value allows to considerably save memory to track the process. To sum up, we define a transition matrix $\mathbf{H}$ taking indices in $\mathbb{Z}_N \cup \{\omega, \delta\}$, defined as follows:

$$h_{ij} = 0\,, \qquad h_{i\delta} = \sum_{k=0}^{\chi(i)-1} \pi_k\,, \qquad h_{i\omega} = \sum_{k=\chi(i)}^{N} \pi_k\,,$$

$$h_{\delta j} = \pi_j\,, \qquad h_{\delta\delta} = 0\,, \qquad h_{\delta\omega} = \pi_N\,,$$

$$h_{\omega j} = \pi_j\,, \qquad h_{\omega\delta} = 0\,, \qquad h_{\omega\omega} = \pi_N\,,$$

where $i$ and $j$ are generic elements of $\mathbb{Z}_N$ and $\pi_j$ is the probability that the channel is in state $j$, as per (7).

We can find the estimated distribution of $\tau_R$, called $\widetilde{\eta}(\tau_R)$, as follows. Given that the HARQ chain visits the set of states $\{\omega, \delta\}$, which happens each time a new packet is transmitted, $\widetilde{\eta}(1)$ is the probability of being in state $\omega$ coming from $\omega$ or $\delta$; $\widetilde{\eta}(2)$ is the probability of being in $\omega$ coming from a state in $\mathbb{Z}_N$; finally, $\widetilde{\eta}(3)$ is the probability of being in $\delta$.

To quantify the goodness in the approximation of this last value, we take the Kullback Leibler divergence [37, p. 18], a well-known measure of the inefficiency in distribution estimations. For the case under study, it quantifies the difference between the true outcome of any HARQ frame and its estimation obtained with the quantized channel. The Kullback Leibler divergence $D(\eta(\tau_R) \parallel \widetilde{\eta}(\tau_R))$ arises as an expected logarithm of the likelihood ratio of the two distributions. As $\tau_R$ can equal 1, 2, or 3, it holds

$$D(\eta(\tau_R) \parallel \widetilde{\eta}(\tau_R)) = \sum_{\tau_R=1}^{3} \eta(\tau_R) \log_2 \frac{\eta(\tau_R)}{\widetilde{\eta}(\tau_R)}. \qquad (19)$$

As in [37], in (19) we conventionally assume $0 \log(0/x) = 0$ and $x \log(x/0) = +\infty$. This allows to deal with degenerate cases, e.g., where the estimate always considers the frame to be acknowledged.

## V. Numerical Evaluation

We use the SNR thresholds derived in [7], [36] for good binary LDPC and Turbo codes ensembles, transmitted over parallel channels with random assignments. We refer the interested reader to these papers for details on the threshold derivation and assumptions. For the analytical framework reported above, the reliable channel region is described by means of

$$\vartheta_1 = \log\left(\frac{\rho}{e^{-c_0} - 1 + \rho}\right)$$

$$\vartheta_2(s_1) = \log\left(\frac{\rho}{e^{-c_0} - 1 + 2\rho - \rho e^{-s_1}}\right)$$

where $\gamma_0$ is the average SNR, $\rho$ is the symbol assignment probability and $c_0$ is the code ensemble noise threshold, that depends on the code ensemble and the code rate.

In the following we compare two different approaches: a uniform probability (UP) SNR quantization method [29], and our proposed CM technique. In particular, for this latter model we utilize the system of equations resulting from (17), solved through standard numerical tools to determine the thresholds $\alpha_1, \alpha_2, \ldots, \alpha_N$. For the former, we use a partitioning rule given by the equiprobability of the states. Also, similarly to the CM approach, we choose the approximation of the reliable region according to a ML criterion. Both approaches are tested for LDPC codes with code rate $1/5$ (labeled in the figures as "LDPC 1/5") and Turbo Codes with code rate $1/3$ ("Turbo 1/3"). Similar results can be obtained for other code ensembles and/or code rates with different threshold functions.

In Fig. 3 we report the value of the quantization error probability on the transmission instant $Q_R$ obtained by both CM and UP approaches, as a function of the number of states $N + 1$, where $N$ is the number of threshold points, with $\gamma_0 = 6$ dB. Compared to the UP approach that does not involve any optimization of $Q_R$, the CM quantization achieves a significant advantage and requires much fewer states. For example, for the LDPC code a very good approximation (less than $0.5\%$ of false positives and false negatives) is achieved with $N = 2$ only, i.e., 3 states in the channel model, where UP requires 10 states to have the same degree of approximation. A similar comparison also holds for the Turbo code: the CM approach obtains with 5 states the same performance as the UP quantization with 18 states. Note also that oscillations are present in the UP curve, which can be explained as due to the lack of code-awareness when choosing the thresholds:

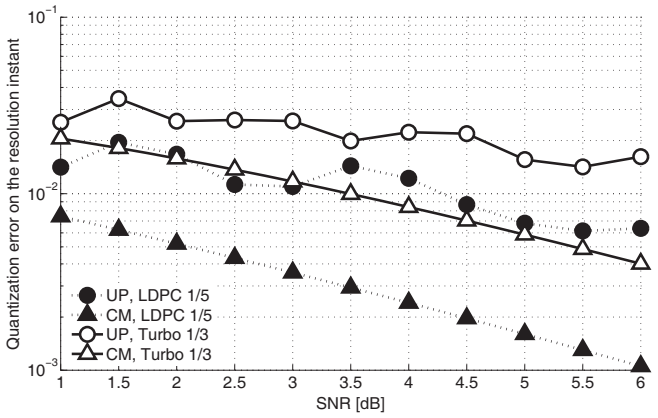



Fig. 5. Quantization error probability on the resolution instant, $Q_R$, of the CM and UP quantization methods versus the average SNR, $R = 1$ bps/Hz, two transmissions, 8 channel states.

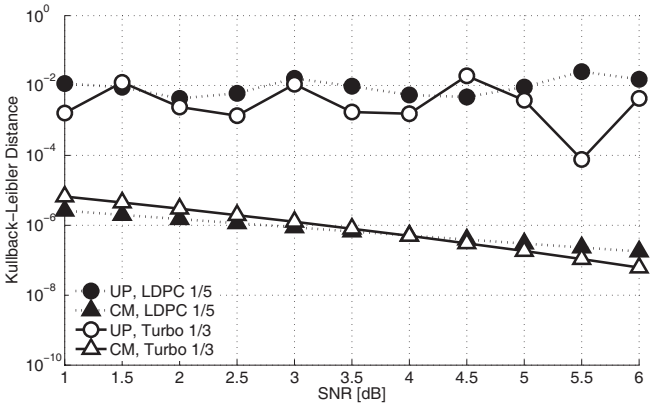



Fig. 6. Kullback Leibler divergence of the distribution of $\tau_R$ for the CM and UP quantization methods versus the average SNR, $R = 1$ bps/Hz, two transmissions, 8 channel states.

whereas in the CM approach the addition of a threshold is always beneficial, as per Lemma 1, in the UP case this is not necessarily true, since a finer partition (i.e., increasing the number of thresholds) does not always correspond to a closer fit of the region.

In Fig. 4 we consider the relative entropy between the true distribution of $\tau_R$, $\eta(\tau_R)$, and the estimated distribution $\widetilde{\eta}(\tau_R)$ obtained in the quantized channel. Differently from $Q_R$ itself, which is simply an indicator of an incorrect approximation, this metric is more closely related to the performance of HARQ. The figure shows that the CM approach is better able to represent HARQ processes from the viewpoint of higher layers (i.e., frame outcome) than UP, and the performance gap is even more relevant than for the simpler quantization error. Especially, observe that, already for a small number of states, the CM approach pushes down the Kullback Leibler divergence to very small values, whereas the UP approach goes to infinity. When the number of states is increased, an oscillatory behavior is still present for the UP quantization, which has the same interpretation as previously discussed. Conversely, the CM approach steadily improves the approximation.

Figs. 5–6 report the same metrics as in Figs. 3–4 varying the average SNR $\gamma_0$ (in dB), for the case $N = 7$ (i.e., 8 channel states). For both figures, note that not only does the UP strategy perform worse than CM in terms of area error and relative entropy, but also it keeps oscillating. Thus, the UP approach does not guarantee an improvement if the SNR is increased, and the error may be significant even for high $\gamma_0$. On the other hand, we notice not only a general better adherence to reality (which enables an improved performance evaluation) obtained by means of our proposed model, but also a steeper descent of the metrics when the channel quality is improved.

To sum up, the code matched quantization technique is shown to offer a channel representation better adhering to HARQ performance, from both viewpoints of low layers (minimum area error) and high layers (significantly better description of metrics related to HARQ frames). For these reasons, our proposed technique can be an extremely useful tool to assess HARQ performance.

## VI. Conclusions

In this paper, we introduced a novel channel quantization method suited for the study of Incremental Redundancy Hybrid ARQ. Our proposal is motivated by the strong research interest around retransmission-based error control schemes and the renewed attention gained by practical coding techniques, such as LDPC and Turbo codes, that show a threshold behavior, i.e., they have error probability asymptotically going to zero if the channel parameters fall within the so-called reliable region.

We used a quantized channel representation to develop a finite-state channel model and to assess the performance of a Stop-and-Wait IR-HARQ scheme. We presented numerical evaluations showing the superior performance in terms of channel representation accuracy, as well as higher layer HARQ metrics, of the proposed quantization with respect to another alternative technique widely used in the literature, i.e., the quantization with uniform probability. Hence, we believe that our proposed methodology can be extremely useful to achieve a compact and accurate channel representation for both analytical and simulation evaluations of HARQ systems.

## Appendix I
## Proof of Lemma 1

*Proof:* For graphical aid in this proof, refer to Fig. 7. We need to prove that the approximation of $\mathcal{U}$ obtained with $N$ thresholds can always be improved by inserting a proper additional threshold. Determine the point $x^*$ such that $x^* = \Theta(x^*)$. If $x^*$ does not coincide with any of the thresholds $\sigma_0, \sigma_1, \ldots, \sigma_N$, the proof is promptly obtained. In fact, for symmetry reasons, $x^*$ must belong to a probability-box with the same indices, say $j$, for column and row, i.e., $x^* \in Y_j \times Y_j$. Recall that $Y_j = [\sigma_j, \sigma_{j+1}]$. Thus, by choosing $x^*$ as the additional threshold, $Y_j$ is further partitioned into $Y' = [\sigma_j, x^*]$ and $Y'' = [x^*, \sigma_{j+1}]$, and correspondingly $Y_j \times Y_j$ is divided into four parts. The ML criterion determines that $Y' \times Y'$ is unresolved and $Y'' \times Y''$ is resolved, so the approximation of $\mathcal{U}$ in any case is improved. On the other hand, assume $x^*$ is already a threshold, say $\sigma_{j^*}$, with $j^* > 0$, and consider $Y_{j^*} \times Y_{j^*-1}$. If this probability-box is resolved, it

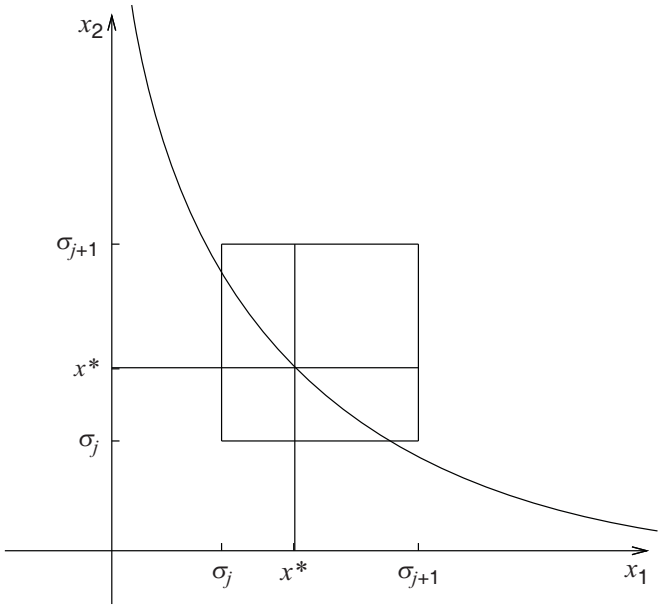



Fig. 7. Reference figure for the proof of Lemma 1.

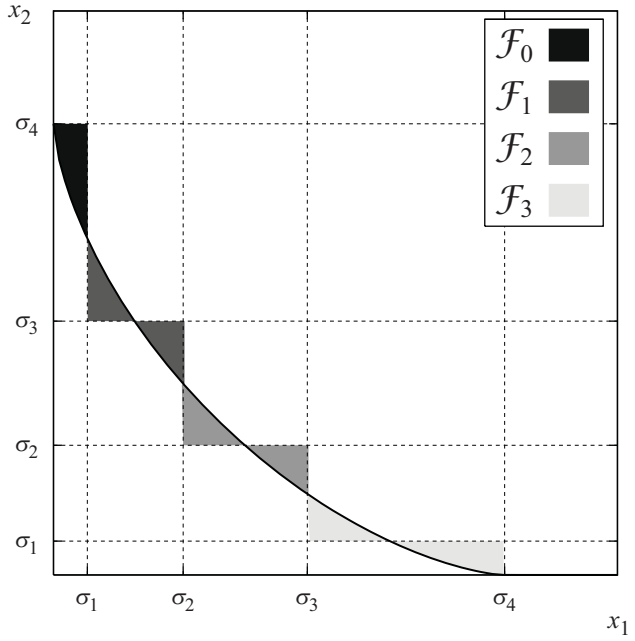



Fig. 8. Thresholds and area error regions of the $i$-approach for $N = 4$.

is easy to see that there exists a positive $\delta$, small enough such that adding $\sigma_{j^*}+\delta$ as a threshold, and calling, according to the ML criterion, $[\sigma_{j^*}, \sigma_{j^*}+\delta]\times Y_{j^*-1}$ as unresolved improves the approximation of $\mathcal{U}$. Analogously, if $Y_{j^*}\times Y_{j^*-1}$ is unresolved, one can make a similar reasoning on $\sigma_{j^*}-\delta$, where again $\delta$ is a sufficiently small positive value. ■

## APPENDIX II
## PROOF OF THEOREM 4

*Proof:* For graphical aid in the proof, refer to Fig. 8 where an application of the i-approach for the case $N = 4$ is plotted. The shaded portion is the error region $\mathcal{W}$. In particular, $\sigma_N = F_\Gamma(\vartheta_1)$, and the other probability thresholds subdivide $\mathcal{W}$ into $N$ parts, called *error subregions*, denoted with $\mathcal{F}_k$, $k = 0, 1, \ldots, N-1$ and formally defined as $\mathcal{F}_k = \mathcal{W}\cap(Y_k\times[0,1])$. They are plotted in Fig. 8 with different shades of grey. Let $\varphi_k$ be the area of the $k$th error subregion $\mathcal{F}_k$.

The error subregions are disjoint so we can evaluate the area of $\mathcal{W}$ as the sum of $\varphi_k$. Notice also that all the points of the first error subregion, $\mathcal{F}_0$, have a value of $x_2$ which is between $\Theta(x_1)$ and $\sigma_N$. For $k > 0$, $\mathcal{F}_k$ comprises two parts, one above and one below the $\Theta(x_1)$ curve.

Formally, we can write:

$$\begin{aligned}\varphi_k &= \int_{\sigma_k}^{\sigma_{k+1}} |\Theta(x_1) - \sigma_{N+1-k}|\,\mathrm{d}x_1 \\ &= \int_{\sigma_k}^{\Theta(\sigma_{N+1-k})} (\Theta(x_1) - \sigma_{N+1-k})\,\mathrm{d}x_1 \\ &+ \int_{\Theta(\sigma_{N+1-k})}^{\sigma_{k+1}} (\sigma_{N+1-k} - \Theta(x_1))\,\mathrm{d}x_1\end{aligned}$$

The symmetry of the curve $\Theta$ implies that $\Theta(\Theta(x)) = x$. By exploiting this property in the relationship $Q_R = \sum_{k=0}^{N}\varphi_k$ one can take the first order derivative with respect to $\sigma_k$, obtaining:

$$\mathrm{d}Q_R/\mathrm{d}\sigma_k = 2\sigma_{N+1-k} - 4\Theta(\sigma_k) + 2\sigma_{N+2-k}\,,$$

where all the resulting terms $\Theta(\Theta(\sigma_k)) - \sigma_k$ are equal to 0. By imposing all derivatives to be equal to 0, (17) is obtained. This is also shown to correspond to a minimum as the second order derivative is $\mathrm{d}^2Q_R/\mathrm{d}\sigma_k^2 = -4(\mathrm{d}\Theta/\mathrm{d}\sigma_k) > 0$. ■

## APPENDIX III
## PROOF OF THEOREM 6

*Proof:* The case $N = 1$ is not meaningful, as the only SNR threshold is placed in $\vartheta_1$. Thus, we need to consider $N > 1$. For the sake of simplicity, we give the proof only for $N = 2$, where $\sigma_2 = F_\Gamma(\vartheta_1)$ and a single threshold is added between 0 and $\sigma_2$. This situation is depicted in Figs. 9 and 10 for the i-approach and the x-approach, respectively. The extension to $N > 2$ is straightforward.

For the sake of exposition, we denote the value of $\sigma_1$ differently for the i-approach and the x-approach. Let us call $\sigma_I$ and $\sigma_X$ the optimal thresholds in the i-approach and in the x-approach, respectively. Analogously, the minimal quantization error probability $Q_R$ obtained in the i-approach and the x-approach is denoted with $Q_I$ and $Q_X$, respectively. As a consequence of (17) and (18), $\Theta(\sigma_I) = \frac{1}{2}(\sigma_I + \sigma_2)$ and $\Theta(\sigma_X) = \sigma_X/2$. It is also possible to prove that $\sigma_X > \sigma_I > \sigma_X/2$. As a result, points $(\sigma_X, \sigma_X)$ and $(\sigma_I, \sigma_I)$ are always above and below the curve $\Theta(x_1)$, respectively, as shown in Figs. 9 and 10.

From Fig. 9, observe that the area marked with a darker shade *minus* the areas marked with a lighter shade equals $\int_0^{\sigma_2}\Theta(x_1)\mathrm{d}x_1 + \sigma_I^2 - 2\sigma_I\sigma_2$. Therefore,

$$Q_I = \int_0^{\sigma_2}\Theta(x_1)\mathrm{d}x_1 + \sigma_I^2 + 2\sigma_I\sigma_2 - 4\int_0^{\sigma_I}\Theta(x_1)\mathrm{d}x_1.$$

Since because of the symmetry of the curve (see Fig. 9)

$$\int_0^{\sigma_I}\Theta(x_1)\mathrm{d}x_1 = \sigma_I(\sigma_I+\sigma_2)/2 + \int_{\frac{\sigma_I+\sigma_2}{2}}^{\sigma_2}\Theta(x_1)\mathrm{d}x_1$$

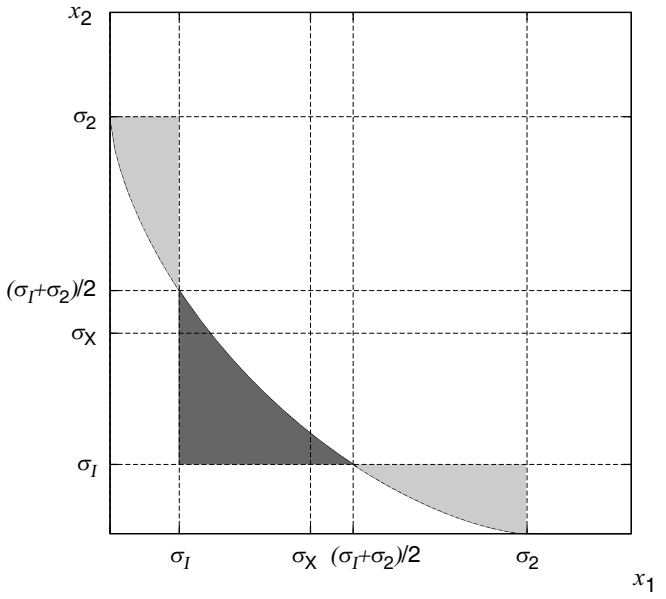



Fig. 9. Single-threshold case of the $i$-approach and resulting area errors.

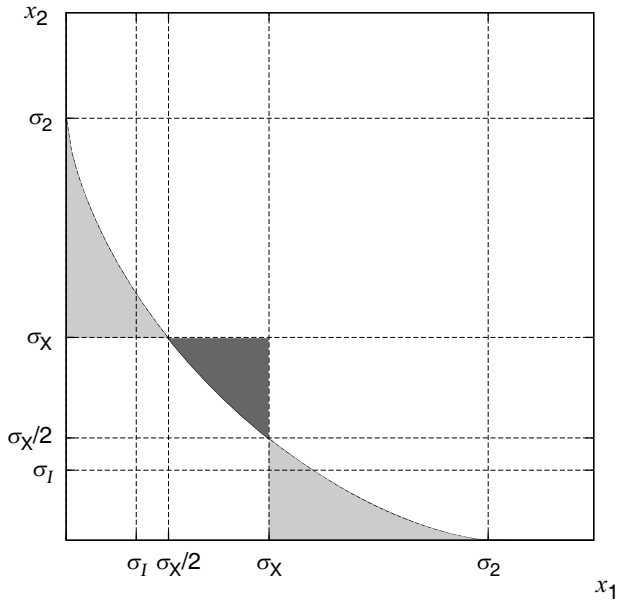



Fig. 10. Single-threshold case of the $x$-approach and resulting area errors.

we have

$$Q_I \;=\; \int_0^{\sigma_2} \Theta(x_1)\mathrm{d}x_1 - \sigma_I^2 - 4\int_{\frac{\sigma_I+\sigma_2}{2}}^{\sigma_2} \Theta(x_1)\mathrm{d}x_1\,. \quad (20)$$

Instead, the quantization error probability of the x-approach can be written as

$$Q_X \;=\; 2\int_{\sigma_X}^{\sigma_2} \Theta(x_1)\mathrm{d}x_1 + \frac{\sigma_X^2}{2} - \int_{\sigma_X/2}^{\sigma_X} \Theta(x_1)\mathrm{d}x_1,$$

where the first and the remaining terms comprise the regions with light grey and dark grey shade, respectively, in Fig. 10.

We can re-arrange this expression by observing that:

$$\int_0^{\sigma_X/2} \Theta(x_1)\mathrm{d}x_1 = \int_{\sigma_X}^{\sigma_2} \Theta(x_1)\mathrm{d}x_1 + \sigma_X\frac{\sigma_X}{2}\,,$$

and therefore

$$Q_X \;=\; \int_0^{\sigma_2} \Theta(x_1)\mathrm{d}x_1 - 2\int_{\sigma_X/2}^{\sigma_X} \Theta(x_1)\mathrm{d}x_1\,. \quad (21)$$

To prove that $Q_I \le Q_X$, we compare (20) and (21). We need to show that

$$\sigma_I^2 + 4\int_{(\sigma_I+\sigma_2)/2}^{\sigma_2} \Theta(x_1)\mathrm{d}x_1 \ge 2\int_{\sigma_X/2}^{\sigma_X} \Theta(x_1)\mathrm{d}x_1. \quad (22)$$

Because of concavity, $\Theta'(\frac{\sigma_I+\sigma_2}{2}) \ge -1$, so that

$$\int_{(\sigma_I+\sigma_2)/2}^{\sigma_2} \Theta(x_1)\mathrm{d}x_1 \ge \sigma_I^2/2.$$

The concavity of the curve also implies that the region below $\Theta$ between $\sigma_X/2$ and $\sigma_X$ is all contained within the trapezoid with vertices $(\sigma_X/2, 0)$, $(\sigma_X, 0)$, $(\sigma_X, \sigma_X/2)$, and $(\sigma_X/2, \sigma_X)$, whose area is $\frac{3}{8}\sigma_X^2$. Therefore, (22) is proved since its left-hand member is not less than $3\sigma_I^2 \ge \frac{3}{4}\sigma_X^2$ which is in turn not less than the right-hand member. ■

## REFERENCES


[1] S. Lin, D. J. Costello, and M. J. Miller, "Automatic-Repeat-reQuest error control schemes," *IEEE Commun. Mag.*, vol. 22, no. 12, pp. 5–17, Dec. 1984.

[2] S. Kallel, "Complementary punctured convolutional (CPC) codes and their applications," *IEEE Trans. Commun.*, vol. 43, no. 6, pp. 2005–2009, June 1995.

[3] I. Sason and S. Shamai, "On improved bounds on the decoding error probability of block codes over interleaved fading channels, with applications to turbo–like codes," *IEEE Trans. Inform. Theory*, vol. 47, no. 6, pp. 2275–2299, Sep. 2001.

[4] D. J. C. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 399–431, Mar. 1999.

[5] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo codes," *IEEE Trans. Commun.*, vol. 44, no. 10, pp. 1261–1271, Oct. 1996.

[6] D. J. C. MacKay and R. M. Neal, "Near Shannon limit performance of low density parity check codes," *IEE Electron. Lett.*, vol. 33, no. 6, pp. 457–458, May 1993.

[7] R. Liu, P. Spasojevic, and E. Soljanin, "Reliable channel regions for good binary codes transmitted over parallel channels," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1405–1424, Apr. 2006.

[8] S. Litsyn and V. Shevelev, "On ensembles of low-density parity-check codes: Asymptotic distance distributions," *IEEE Trans. Inform. Theory*, vol. 48, no. 4, pp. 887–907, Apr. 2002.

[9] I. Sason, E. Telatar, and R. Urbanke, "On the asymptotic input–output weight distributions and thresholds of convolutional and turbo–like encoders," *IEEE Trans. Inform. Theory*, vol. 48, no. 12, pp. 3052–3061, Dec. 2002.

[10] H. Jin and R. J. McEliece, "Coding theorems for turbo code ensembles," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1451–1461, June 2002.

[11] I. Sason, S. Shamai, and D. Divsalar, "Tight exponential upper bounds on the ML decoding error probability of block codes over fully interleaved fading channels," *IEEE Trans. Commun.*, vol. 51, no. 8, pp. 1296–1305, Aug. 2003.

[12] I. Sason and S. Shamai, "Improved upper bounds on the ML decoding error probability of parallel and serial concatenated turbo codes via their ensemble distance spectrum," *IEEE Trans. Inform. Theory*, vol. 46, no. 1, pp. 24–47, Jan. 2000.

[13] ——, *Performance Analysis of Linear Codes under Maximum-Likelihood Decoding: A Tutorial*, ser. Foundations and Trends in Communications and Information Theory. Delft, the Netherlands: NOW Publishers, 2006, vol. 3, no. 1–2.

[14] L. Bazzi, T. J. Richardson, and R. Urbanke, "Exact thresholds and optimal codes for the binary–symmetric channel and Gallager's decoding

algorithm A," *IEEE Trans. Inform. Theory*, vol. 50, no. 9, pp. 2010–2021, Sep. 2004.

[15] T. J. Richardson and R. Urbanke, "The capacity of low–density parity–check codes under message–passing decoding," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.

[16] A. W. Eckford, F. R. Kschischang, and S. Pasupathy, "Analysis of low–density parity–check codes for the Gilbert–Elliott channel," *IEEE Trans. Inform. Theory*, vol. 51, no. 11, pp. 3872–3889, Nov. 2005.

[17] ——, "On designing good LDPC codes for Markov channels," *IEEE Trans. Inform. Theory*, vol. 53, no. 1, pp. 5–21, Jan. 2007.

[18] O. F. Acikel and W. E. Ryan, "Punctured turbo-codes for BPSK/QPSK channels," *IEEE Trans. Commun.*, vol. 47, no. 9, pp. 1315–1323, Sep. 1999.

[19] B. Shen, A. Patapoutian, and P. A. McEwen, "Punctured recursive convolutional encoders and their applications in turbo codes," *IEEE Trans. Inform. Theory*, vol. 47, no. 6, pp. 2300–2320, Sep. 2001.

[20] F. Babich, G. Montorsi, and F. Vatta, "Some notes on rate-compatible puntured turbo codes (RCPTC) design," *IEEE Trans. Commun.*, vol. 52, no. 5, pp. 1616–1625, May 2004.

[21] K. R. Narayanan and G. L. Stüber, "A novel ARQ technique using the turbo coding principle," *IEEE Commun. Lett.*, vol. 1, pp. 49–51, Mar. 1997.

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

[23] I. Sason and I. Goldenberg, "Coding for parallel channels: Gallager bounds for binary linear codes with applications to turbo–like codes," *IEEE Trans. Inform. Theory*, vol. 53, no. 7, pp. 2394–2428, July 2007.

[24] Q. Zhang, T. F. Wong, and S. Lehnert, "Performance of a type–II hybrid ARQ protocol in slotted DS–SSMA packet radio systems," *IEEE Trans. Commun.*, vol. 47, no. 2, pp. 281–290, Feb. 1999.

[25] S. Kallel, R. Link, and S. Bakhtiyari, "Throughput performance of memory ARQ scheme," *IEEE Trans. Veh. Technol.*, vol. 48, no. 3, pp. 891–899, May 1999.

[26] 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.

[27] Q. Zhang and S. A. Kassam, "Hybrid ARQ with selective combining for fading channel," *IEEE J. Select. Areas Commun.*, vol. 17, no. 5, pp. 867–880, May 1999.

[28] C. H. C. Leung, Y. Kikumoto, and S. A. Sorensen, "The throughput efficiency of the go-back-N ARQ scheme under Markov and related error structures," *IEEE Trans. Commun.*, vol. 36, no. 2, pp. 231–234, Feb. 1988.

[29] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[30] C. C. Tan and N. C. Beaulieu, "On first-order Markov modeling for the Rayleigh fading channel," *IEEE Trans. Commun.*, vol. 48, no. 12, pp. 2032–2040, Dec. 2000.

[31] M. Rossi, L. Badia, and M. Zorzi, "SR ARQ delay statistics on N-state Markov channels with non-instantaneous feedback," *IEEE Trans. Wireless Commun.*, vol. 5, no. 6, pp. 1526–1536, June 2006.

[32] A. Guillen i Fabregas and G. Caire, "Coded modulation in the block-fading channel: Coding theorems and code construction," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 91–114, Jan. 2006.

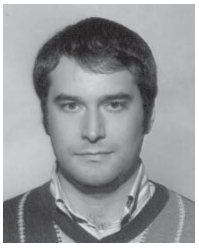
[33] W. Vijacksungsithi and K. Winick, "Joint channel-state estimation and decoding of low-density parity-check codes on the two-state noise-less/useless BSC block interference channel," *IEEE Trans. Commun.*, vol. 53, no. 4, pp. 612–622, Apr. 2005.

[34] D. Garg and F. Adachi, "Rate compatible punctured turbo-coded hybrid ARQ for OFDM in a frequency selective fading channel," in *Proc. IEEE VTC - Spring*, Apr. 2003, vol. 4, pp. 2725–2729.

[35] L. Badia, M. Rossi, and M. Zorzi, "SR ARQ packet delay statistics on Markov channels in the presence of variable arrival rate," *IEEE Trans. Wireless Commun.*, vol. 5, no. 7, pp. 1639–1644, July 2006.

[36] E. Soljanin, N. Varnica, and P. Whiting, "Punctured vs rateless codes for hybrid ARQ," in *Proc. IEEE Information Theory Workshop*, 2006, pp. 155–159.

[37] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, Inc., 1991.



**Leonardo Badia** (S'02-M'04) was born in Ferrara, Italy, in 1977. He received the Laurea Degree (with honors) in Electrical Engineering and the Ph.D. in Information Engineering from the University of Ferrara, Italy, in 2000 and 2004, respectively. During 2002 and 2003 he was on leave at the Radio System Technology Labs (now Wireless@KTH), Royal Institute of Technology of Stockholm, Sweden. After having been with the Engineering Department of the Università di Ferrara, Italy, he joined in 2006 the "Institutions Markets Technologies" (IMT) Lucca Institute for Advanced Studies, in Lucca, Italy, where he is currently a Research Fellow. He also collaborates with DEI, University of Padova, Italy. His research interests include energy efficient Ad Hoc Networks, transmission protocol modeling, Admission Control and economic modeling of Radio Resource Management for Wireless Networks. Dr. Badia serves also as a reviewer for several periodicals in the communication area.

**Marco Levorato** (S'06) was born in Venice on March 18th, 1980. He obtained both the BE (Electronics and Telecommunications Engineering) and the ME (Telecommunications Engineering) *summa cum laude* from the University of Ferrara (Italy) in 2002 and 2005, respectively. During 2005 he held a fellowship at the University of Padova (Italy), and from January 2006 he has been a Ph.D. student in Information Engineering at the University of Padova under the supervision of Prof. Michele Zorzi. During 2008 he was on leave at the University of Southern California, Los Angeles. His research interests include cooperative communications, design of ad hoc networks with multiuser detection, and analysis of Hybrid ARQ techniques.

**Michele Zorzi** (S'89-M'95-SM'98-F'07) was born in Venice, Italy, in 1966. He received the Laurea degree and the Ph.D. degree in Electrical Engineering from the University of Padova, Italy, in 1990 and 1994, respectively. During the Academic Year 1992/93, he was on leave at the University of California, San Diego (UCSD), attending graduate courses and doing research on multiple access in mobile radio networks. In 1993, he joined the faculty of the Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy. After spending three years with the Center for Wireless Communications at UCSD, in 1998 he joined the School of Engineering of the University of Ferrara, Italy, and in 2003 joined the Department of Information Engineering of the University of Padova, Italy, where he is currently a Professor. His present research interests include performance evaluation in mobile communications systems, random access in mobile radio networks, ad hoc and sensor networks, and energy constrained communications protocols.

Dr. Zorzi was the Editor-In-Chief of *IEEE Wireless Communications Magazine* from 2003 to 2005, is currently the Editor-In-Chief of the IEEE TRANSACTIONS ON COMMUNICATIONS, and serves on the Steering Committee of the IEEE TRANSACTIONS ON MOBILE COMPUTING, and on the Editorial Boards of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the *Wiley Journal of Wireless Communications and Mobile Computing* and the *ACM/URSI/Kluwer Journal of Wireless Networks*. He was also guest editor for special issues in *IEEE Personal Communications Magazine* (Energy Management in Personal Communications Systems) and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (Multi-media Network Radios).