An Improved Channel Quantization Method for Performance Evaluation of Incremental Redundancy HARQ Based on Reliable Channel Regions

Leonardo Badia*†, Marco Levorato†, Michele Zorzi†

* IMT Lucca Institute for Advanced Studies, Piazza San Ponziano 6, 55100 Lucca, Italy

† Department of Information Engineering, University of Padova, Via Gradenigo 6/B, 35131 Padova, Italy

email: *l.badia@imtlucca.it*, {*levorato,zorzi*}*@dei.unipd.it*

***Abstract*—In this paper, we investigate useful channel representations for Incremental Redundancy Hybrid ARQ. When these techniques are employed, parts of the codeword are transmitted 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, a methodology which has been often used in the past for the study of the performance of ARQ protocols, and we propose a novel method to derive efficient channel partitioning rules, i.e., a code-matched quantization of the channel state. We present results showing the performance of our proposed method and its capability of representing the channel in a compact and accurate way.**

***Index Terms*—Channel modeling, hybrid automatic repeat request, Markov processes, error correction techniques, channel coding.**

I. Introduction

Incremental Redundancy (IR) implementations of Hybrid Automatic Repeat reQuest (HARQ) schemes have been widely employed for error control in noisy channels. An assumption commonly made about the codes used in IR-HARQ schemes is that they are characterized by a threshold behavior, i.e., the error probability can be regarded as zero if the channel realization is inside a region (the so called *reliable region*), and as one otherwise [1]. Although this is an abstraction, it works reasonably well to approximate the performance of practical codes, e.g., Turbo codes [2] and Low–Density Parity–Check (LDPC) codes [3].

Even under this simplification, it is important to know whether the channel parameters fall within the reliable region or not. Thus, an accurate representation of the channel is still key in evaluating the performance of HARQ policies. On the other hand, IR-HARQ systems are usually employed in time–varying channel contexts, among which a very important example is represented by wireless fading channels. In such cases, tracing the channel evolution may require to store an excessive amount of information; for this reason, it is also 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 can be represented by a Finite-State Markov Channel (FSMC) description, employing a proper channel quantization rule, i.e., representing the channel state with an index spanning over a discrete set. However, the FSMC approaches presented in the literature [4], [5] are mostly based on physical layer characteristics, rather than higher layer performance aspects. Instead, in the present paper we propose to utilize a *Code-Matched* (CM) channel quantization, meaning that we explicitly take into account the coding performance in partitioning the channel. Our quantization directly aims at giving an efficient approximation, with a given number of states, of the reliable region, so as to properly characterize the performance of the IR-HARQ scheme.

In a sense, the contribution given by the present paper does not strictly depend on the channel representation through a Markov Chain, which requires several assumptions such as exponential sojourn time in the states. The rationale proposed by our investigation might be applied identically to other models such as Hidden Markov Models (HMM) or Semi-Markov Models (SMM), which seek to improve the accuracy in the channel representation [6]. The choice of applying a code-matched channel representation to FSMC is due only to the widespread usage of this model, but it could be identically applied to HMM or SMM as well.

In particular, in this paper we instantiate our proposal of utilizing a CM approach to analyze, by means of a Markov model, an IR-HARQ system based on a Stop-and-Wait policy. We give a detailed analytical characterization of the case where the HARQ system adopts a two-transmission limit. The extension to cases with a higher number of maximum retransmissions can be done along the same lines. We also present numerical results to quantify the goodness of our proposed approach in assessing IR-HARQ performance, together with existing channel quantization techniques (e.g., equiprobable states [4]). We numerically evaluate the performance bringing examples of different codes, namely LDPC and Turbo, which can be used in the IR-HARQ scheme.

These evaluations show that, compared to other techniques, the proposed 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 the problem statement is provided. Here, we describe the IR-HARQ mechanism assumed in the analysis and we introduce the concept of reliable region of the code, which is key in deriving the code matched channel quantization. Section III focuses on the two-transmission case, for which we state the main analytical results about the code matched channel quantization reported in detail in the appendices. In Section IV 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 representations. Finally, we draw the conclusions in Section V.

II. System Model and Problem Statement

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 time in a single HARQ packet. For this reason, in the following we will utilize the terms *frame* and *codeword* interchangeably, and similarly for *packet* and *fragment*. In order to keep the analysis simple, we assume 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 [7].

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 message 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 characteristic of IR-HARQ is that a NACK does trigger a retransmission, but differently from other retransmission-based techniques, a physically different packet (though still associated with the same information frame) is transmitted. Hence, we adopt a slight terminology abuse by speaking, when this event happens, of *frame retransmission*.

For the sake of simplicity, 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 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.

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 in this sense a binary (i.e., threshold-wise) behavior. However, if F is increased to large values, an exact description of this process can become cumbersome since it possibly includes the evaluation of F-dimensional thresholds.

One important case of application of HARQ is for obtaining reliable data transmission over a wireless fading channel. In the following we refer primarily to this scenario, even though the same rationale is directly applicable to other kinds of noisy channels. The outcome of the transmission over a radio channel depends on the Signal-to-Noise Ratio (SNR) at the receiver.

Due to their good trade-off between accuracy and complexity, FSMC models have gained foothold in both analytical and simulation frameworks. Such techniques are based on partitioning the possible received SNR values into a given set of intervals. However, the most common approaches presented in the literature [4], [5], perform such a quantization according to physical layer aspects, such as the equiprobability of the intervals. Our proposal is to employ a quantization *matched* to the channel/code characteristics. This means that we seek for a set of SNR intervals which optimally describe the decoding process of a sequence of packets in terms of accuracy of the acknowledgement/not acknowledgement decision, based on a Maximum Likelihood (ML) criterion. This approach allows us to decrease the complexity of the HARQ description and enables a fully analytical evaluation.

We remark that the FSMC approach is not perfect. In fact, it requires the key assumption that the sojourn time in a state is exponentially distributed, which may not be true in practice [6]. Moreover, as shown in [8], the FSMC does not match perfectly the statistics of the real process, and there is a gap which can not be filled regardless of how many states are used. To cope with these inefficiencies, HMM or SMM can be used. We stress that our rationale can be extended to these models as well. As a first step towards a joint Markov formulation of the channel/protocol behavior, in this paper we consider a block fading model, where the channel conditions are constant during the transmission of an HARQ packet, and independent across different transmissions. Extension to the case of fading channels with memory is currently being developed.

We refer to the codeword fragment received at the kth transmission as w_k, $k = 1, 2, \ldots, F$. Under the block flat fading assumption, each codeword fragment is characterized by a single received SNR coefficient, which is denoted for the

kth fragment with $s_k \in \mathbb{R}_+$. Since we are interested in determining a statistical model of the channel, we take a probability mapping of the SNR values, e.g., if γ is the random variable describing the SNR, we translate any SNR value $s_k \in \mathbb{R}_+$ into a value $q_k \in [0, 1]$ such that $q_k = q(s_k) = \text{Prob}\{\gamma \leq s_k\}$. The exact mapping function q depends on the statistics of the channel, but it is always an increasing (and therefore invertible) function of s_k.

At the kth transmission, the receiver bases the decoding of the codewords on all fragments received up to w_k. We formally define, to this end, the *reliable channel probability region* $\mathcal{R}(k) \subseteq [0, 1]^k$, which contains the k-tuples of the q-values of channel SNR 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 $(q_1, q_2, \ldots, q_k) \in \mathcal{R}(k)$.

The exact specification of $\mathcal{R}(k)$ is determined by the used code, the decoding algorithm and the codeword fragments construction. We will derive our model under the hypotheses that $\mathcal{R}(k)$ is connected, convex and symmetric with respect to permutations of coordinates, i.e., for all i and j, $1 \leq i, j \leq k$, $(q_1, \ldots, q_i, \ldots, q_j, \ldots, q_k) \in \mathcal{R}(k)$ implies that $(q_1, \ldots, q_j, \ldots, q_i, \ldots, q_k) \in \mathcal{R}(k)$ as well. These assumptions hold true for the specific choices used later in the results' section, which represent realistic scenarios, and are reasonable for most practical cases.

Note also that it is always verified that $(q_1, \ldots, q_{k-1}, q_k) \in \mathcal{R}(k)$ and $q'_k > q_k$ imply that $(q_1, \ldots, q_{k-1}, q'_k) \in \mathcal{R}(k)$. Thanks to this property, we can use a representation of $\mathcal{R}(k)$ through a threshold function $\vartheta_k : [0, 1]^{k-1} \to [0, 1]$, defined as follows:

$$\vartheta_k(\mathbf{q}^{(k-1)}) = \inf\{q_k : (q_1, \ldots, q_{k-1}, q_k) \in \mathcal{R}(k)\}, \qquad (1)$$

where $\mathbf{q}^{(k-1)} = (q_1, q_2, \ldots, q_{k-1})$. In other words, the edge of $\mathcal{R}(k)$ is the curve identified by $\vartheta_k(\mathbf{q}^{(k-1)})$ in $[0, 1]^k$. Note that, when one transmission is considered, this curve degenerates to a single point $\vartheta_1 \in [0, 1]$, which is the value of q_1 associated with the (constant) SNR threshold to obtain correct codeword delivery with a single fragment, and the reliable channel region $\mathcal{R}(1)$ corresponds to the interval $[\vartheta_1, 1]$.

Additionally, observe that $(q_1, \ldots, q_k) \in \mathcal{R}(k)$ also implies that $(q_1, \ldots, q_k, q_{k+1}) \in \mathcal{R}(k+1)$ for all $q_{k+1} \in [0, 1]$, since the fragments $w_1, w_2, \ldots, w_k$ were already sufficient to decode the codeword. Hence, the transmission of a codeword can be dismissed after a success is achieved. Note that this holds under the assumption of perfect feedback. Therefore, in the system under investigation, after the reception of a fragment w_k the receiver is able to decode the packet if q_k is above the threshold $\vartheta_k(\mathbf{q}^{(k-1)})$. In this case, no further fragment transmission is required (for this reason, the case where $\mathbf{q}^{(k)} \in \mathcal{R}(k)$ *and* $\mathbf{q}^{(k+1)} \in \mathcal{R}(k+1)$ never occurs in practice, but is considered only for completeness). Otherwise, another fragment is requested, which would be received as fragment w_{k+1} and will be compared with threshold $\vartheta_{k+1}(\mathbf{q}^{(k)})$ and so on.

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 $\mathbf{q}^{(k)}$ must be kept trace of. However, high complexity and memory requirements are implied to accurately track the evolution of a vector of continuous variables. For this reason, it is meaningful to consider a quantization of the SNR to enable a finite-state representation of the channel, where each state represents an interval of SNR values, and which can be used in a FSMC context.

Thus, we partition $[0, 1]$ into $N + 2$ non overlapping[1] adjacent regions $I_0, I_1, \ldots, I_N, I_{N+1}$. The purpose of this partition is to describe 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 the exact SNR values are no longer known, but only which region the SNR falls within. In fact, according to this representation, any sequence of q-values $\mathbf{q}^{(k)}$ (which, in turn, determines a sequence of SNR values $\mathbf{s}^{(k)}$) is described with a sequence of discrete values $\mathbf{d}^{(k)} = (d_1, d_2, \ldots, d_k) \in \{0, 1, \ldots, N, N+1\}^k$.

The k-tuple $\mathbf{d}^{(k)}$ indicates that for the ith received SNR, $s_i \in q^{-1}(I_{d_i})$, for all $i = 1, 2, \ldots, k$. By checking the relative placement of the region $\mathcal{I}(\mathbf{d}^{(k)}) = I_{d_1} \times I_{d_2} \times \cdots \times I_{d_k} \subseteq [0, 1]^k$, which is a hyper-parallelepiped in $[0, 1]^k$, and $\mathcal{R}(k)$, one can infer, in an approximate manner, whether the reception of the fragments $w_1, w_2, \ldots, w_k$ allows the frame to be acknowledged or not. In general, $\mathcal{I}(\mathbf{d}^{(k)})$ may contain both points belonging to and outside $\mathcal{R}(k)$. Hence, in the quantized channel every region is to be called as corresponding to frame resolution or not according to a Maximum Likelihood (ML) criterion, which means to check whether $\mathcal{I}(\mathbf{d}^{(k)}) \cap \mathcal{R}(k)$ has a larger hyper-volume than $\mathcal{I}(\mathbf{d}^{(k)}) \setminus \mathcal{R}(k)$ (in which case $\mathcal{I}(\mathbf{d}^{(k)})$ is considered as a "resolved" region) or vice versa.

Our objective is to reduce the quantization errors introduced by this representation, i.e., the probability that k-tuple $\mathbf{s}^{(k)}$ is an erroneous SNR sequence which corresponds to a discrete k-tuple $\mathbf{d}^{(k)}$ which determines a "resolved" region, or conversely that $\mathbf{s}^{(k)}$ implies resolution but the ML criterion for $\mathbf{d}^{(k)}$ gives retransmission.

Although the framework has been outlined for the general multi-dimensional case up to this point, for ease of analysis and explanation in the next section we specifically focus on the case of a maximum number of two transmissions ($F = 2$), where a frame is discarded after a single unsuccessful retransmission (i.e., both the first transmission and the subsequent retransmission receive a NACK in response).

III. Channel Chain Construction (Two Transmissions)

If two transmissions are considered, region $\mathcal{R}(2)$ is the portion of $[0, 1]^2$ that lies above the curve $q_2 = \vartheta_2(q_1)$. A sample curve is plotted in Fig. 1. Even though the shape of this region may be different according to the code used, the curve $q_2 = \vartheta_2(q_1)$ is always a non-increasing function. We assume that the curve is also concave and symmetric, which is true in many real cases.

[1] They can be thought as closed intervals, since the set of extreme points has zero measure. Also, we take $N + 2$ intervals since a meaningful partition can not have fewer than 2 intervals.

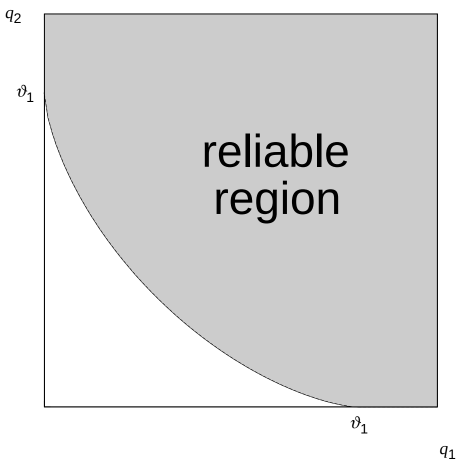

Fig. 1. A sample curve in $\mathbb{R}^2$ and the reliable region $\mathcal{R}(2)$.

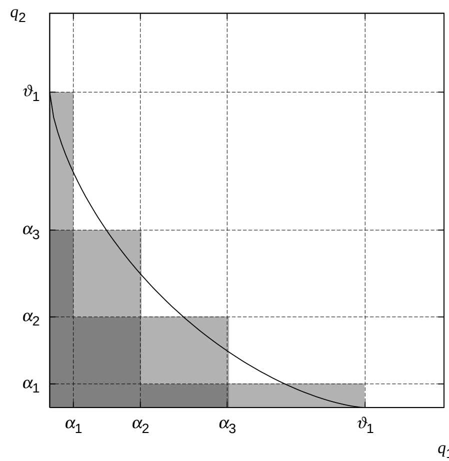

Fig. 2. A graphical comparison of the i-approach and the x-approach.

To find a suitable partition for the 2 transmission case, we therefore proceed as follows. We need to define $N+1$ threshold values $\alpha_1, \alpha_2, \ldots, \alpha_N, \alpha_{N+1}$ which identify the intervals $I_k = [\alpha_k, \alpha_{k+1}]$. For consistency, we conventionally take $\alpha_0 = 0$ and $\alpha_{N+2} = 1$. As explained in the previous section, the optimal partitioning of region $\mathcal{R}(1)$ always corresponds to choosing a single threshold point ϑ_1. Thus, if one threshold value is to be chosen, which happens when $N = 0$, we simply take $\alpha_{N+1} = \vartheta_1$. This choice, which is optimal for a single transmission, will be used for any partitioning of the two-dimensional space $[0, 1]^2$ also.

If $N > 0$, which means that we have additional threshold values to place, we can put them between 0 and ϑ_1; in fact, the region where $q_1 \geq \vartheta_1$ corresponds to a frame resolution (after the first packet transmission), and so does the region where $q_1 < \vartheta_1$ and $q_2 \geq \vartheta_1$ (though this time the frame is resolved after the second packet transmission). Thus, all the remaining N thresholds, i.e., $\alpha_1, \alpha_2, \ldots, \alpha_N$ must be put between 0 and ϑ_1.

There are two possible general strategies to place the α's. In the first one, which in the following we will refer to as *internal approach*, or i-approach for short, the thresholds are placed so that in the two-dimensional rectangular region $I_j \times I_k$ the frame is assumed to be acknowledged if $j + k > N$, and not acknowledged if $j + k \leq N$. In the second strategy, referred to as *external approach*, or x-approach, in $I_j \times I_k$ the frame is assumed to be acknowledged if $j + k \geq N$, and not acknowledged if $j + k < N$. In Fig. 2 we plot a graphical comparison of these two approaches, to show their difference. The i-approach corresponds to approximating the region $\mathcal{R}(2)$ with the white area only, whereas the x-approach considers both white and grey boxes as part of the reliable region.

For both approaches, we want to choose the thresholds so as to minimize the so-called *area error*, i.e., the parts of those regions which are classified as correct but fall outside $\mathcal{R}(2)$ or the parts of those regions which are considered as erroneous even though they are within $\mathcal{R}(2)$. This corresponds to minimizing the probability of wrong decision by the quantized channel, i.e., both false positive and false negative (the continuous channel is in error but the quantized channel calls the frame as correct or vice versa). These two types of error are accounted for with the same weight, but it would possible to extend the analysis to consider different weights as well.

One of the main findings of our investigations is that for both approaches the values of the optimal thresholds follow a general expression, which can be derived in closed-form, for any curve $\vartheta_2(q_1)$. The final expression contains a reference to $\vartheta_2(q_1)$, however the numerical solution of this condition is rather simple since ϑ_2 is a decreasing function and therefore the optimal thresholds $\alpha_1, \alpha_2, \ldots, \alpha_N$ can be directly found via simple numerical methods. Moreover, the uniqueness of the solution is guaranteed.

In the appendices, we prove that:

- Appendix A: the optimal thresholds of the i-approach always satisfy:
$$\vartheta_2(\alpha_i) = \frac{\alpha_{N+2-i} + \alpha_{N+1-i}}{2}, \quad \text{for } i = 1, \ldots, N \quad (2)$$
This set of relationships can be seen as a system of N equations with N unknowns, which, due to the monotonicity of the ϑ_2 function, always admits a unique solution.
- Appendix B: the optimal thresholds of the i-approach always offer a better solution than the optimal thresholds of the x-approach.

Thanks to these theoretical findings, 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. In the next section, we will evaluate the effectiveness of this approach against traditional partitioning techniques, such as the uniform-probability quantization of the SNR.

IV. Performance Evaluation

We present numerical results to quantify the goodness of our proposed approach in assessing IR-HARQ performance. According to the considerations previously made, we focus on a case where $F = 2$, i.e., up to two fragments can be transmitted per codeword, which is thus discarded after the reception of two subsequent NACKs.

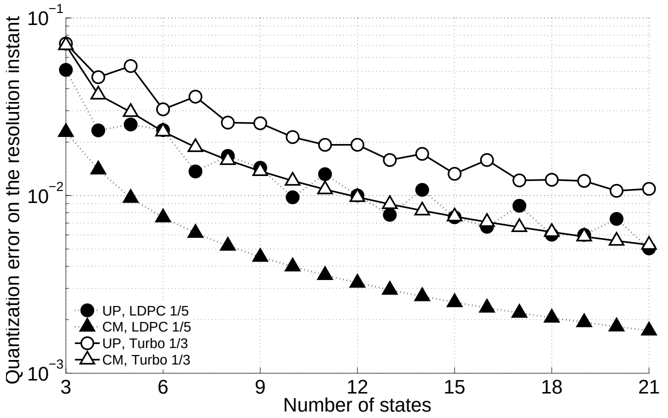

Fig. 3. Area error of the CM and UP quantization methods versus the number of channel states, $R = 2.6$ bps/Hz, $\gamma_0 = 10$ dB, two transmissions.

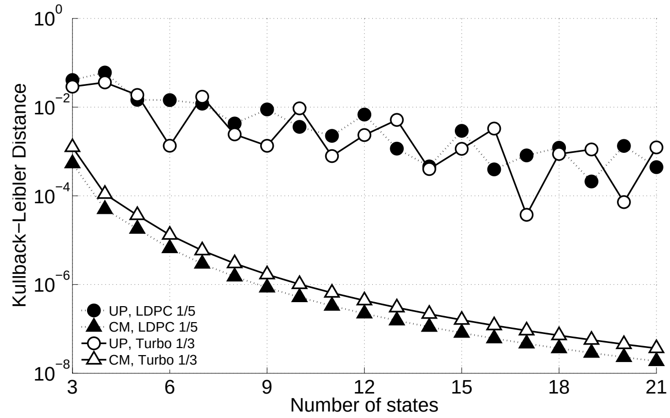

Fig. 4. Relative entropy of the CM and UP quantization methods versus the number of channel states, $R = 2.6$ bps/Hz, $\gamma_0 = 10$ dB, two transmissions.

We use the SNR thresholds derived in [9], [10] 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 thresholds derivation and assumptions. For the analytical framework reported above, the reliable channel region is described by means of

$$\begin{aligned} \vartheta_1 &= 1 - \left(\frac{\rho}{e^{-c_0} + 1 - \rho} \right)^{1/\gamma_0} \\ \vartheta_2(q_1) &= 1 - \left(\frac{\rho}{e^{-c_0} + 1 - \rho(1-q_1)^{-\gamma_0}} \right)^{1/\gamma_0} \end{aligned}$$

where γ_0 is the average SNR, ρ 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 [4], and our proposed CM technique which best fits the thresholds on the reliable channel region. In particular, for this latter model we utilize the system of equations resulting from (2), solved through standard numerical tools to determine the thresholds $\alpha_1, \alpha_2, \ldots, \alpha_N$. Remember that we always consider the number of states to be $N+2$, so the number of thresholds placed by our approach between 0 and ϑ_1 corresponds to the x-axis value decreased by 2. 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 area error (that corresponds to the probability that the actual channel and its quantized version correspond to different transmission outcomes) obtained by both CM and UP approaches. Since the UP approach does not involve any optimization of the area error, we achieve a significant advantage in this sense. Note also that the oscillations in the behavior of the UP curve can be explained by considering that we adopt a ML criterion on the two-dimensional regions $I_j \times I_k$. Whereas in the CM approach the addition of a threshold is always beneficial, since it permits to better approximate the reliable region, 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. Quantitatively, our model permits to utilize much fewer states than UP. For example, for the LDPC code a very good approximation (less than 1% of false positives and false negatives) is achieved with $N = 3$ only, i.e., 5 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 the same performance of the UP quantization with 40% fewer states (e.g., 7 states instead of 12).

In Fig. 4 we consider a metric directly describing HARQ performance, i.e., the number of erroneous transmissions per packet. This variable can equal 0, 1 or 2 (in the case of frame discarding) and, for the analyzed scenario, can be seen as representative also of other HARQ metrics, e.g., throughput. In the quantized channel, the evaluation of this random variable is possibly approximate, and we can compute the relative entropy between the true distribution $p(N_{tx})$ and the estimated distribution $\widetilde{p}(N_{tx})$. This corresponds to the Kullback Leibler divergence (see [11, p. 18]), which is a well-known measure of the inefficiency in distribution estimations.

The Kullback Leibler distance arises as an expected logarithm of the likelihood ratio of the two distributions:

$$D(p(N_{tx}) \parallel \widetilde{p}(N_{tx})) = \sum_{N_{tx}=0}^{2} p(N_{tx}) \log_2 \frac{p(N_{tx})}{\widetilde{p}(N_{tx})}. \tag{3}$$

In the above definition, as in [11], we conventionally assume that when certain probability terms go to zero, we adopt the interpretations $0 \log(0/x) = 0$ and $x \log(x/0) = \infty$. This is necessary to deal with degenerate cases, e.g., where the estimator always considers the frame to be acknowledged.

Fig. 4 shows that our proposed approach is better able to represent higher layer HARQ processes: the CM approach can push down the Kullback Leibler distance to very small values, whereas the conventional quantization strategy obtains a significant relative entropy even for a large number of thresholds, and an oscillatory behavior is still present (which has the same interpretation as previously discussed).

Figs. 5 and 6 report the same curves obtained by varying

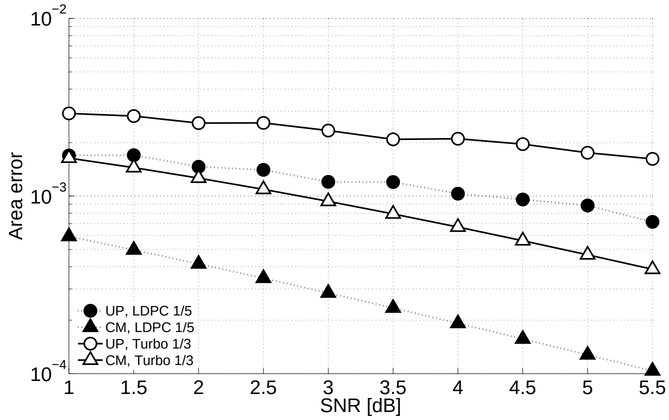

Fig. 5. Area error of the CM and UP quantization methods versus the average SNR, $R = 1$ bps/Hz, two transmissions, 6 channel states.

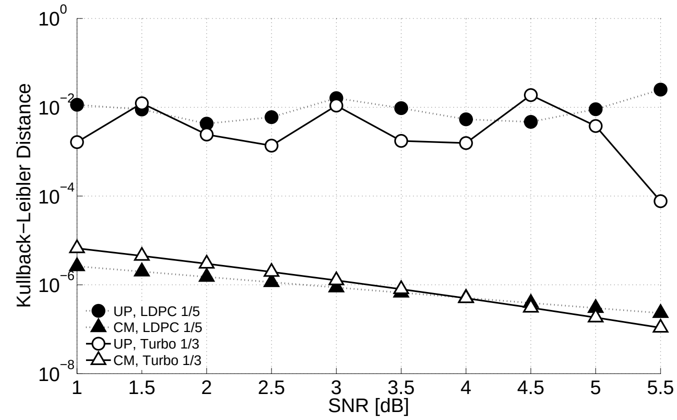

Fig. 6. Relative entropy of the CM and UP quantization methods versus the average SNR, $R = 1$ bps/Hz, two transmissions, 6 channel states.

the average SNR γ_0 (in dB), for the case $N = 4$ (i.e., 6 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 γ_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.

V. Conclusions

In this paper, we proposed a novel channel quantization method especially useful for Incremental Redundancy Hybrid ARQ error control schemes. Motivated by the renewed attention gained by practical coding techniques that show a threshold behavior (e.g., LDPC and Turbo codes), we focused on coding performance models where the error probability asymptotically vanishes if the channel parameters fall within the so-called reliable region, i.e., a given set of channel parameters.

Next, the quantized channel representation is used 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 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 A
Optimal thresholds in the i-approach

In this appendix, we will prove that the optimal thresholds of the i-approach satisfy the condition in (2). For graphical aid, refer to Fig. 7 where an application of the i-approach for the case $N = 3$ is plotted.

Denote the error region (shaded in Fig. 7) with $\mathcal{E}$ and its area with E. From the partition of $[0, \vartheta_1]$ made by the α-thresholds $\alpha_1, \alpha_2, \dots, \alpha_N$, a subdivision of $\mathcal{E}$ can be made into $N+1$ parts, called *error subregions*, denoted with $\mathcal{F}_k$, $k = 0, 1, \dots, N$ and formally defined as $\mathcal{F}_k = \mathcal{E} \cap (I_k \times [0, 1])$. They are plotted in Fig. 7 with different shades of grey. Let φ_k be the area of the kth error subregion $\mathcal{F}_k$.

The error subregions are disjoint so we can evaluate the area of the error region as the sum of the areas of all error subregions. Notice also that all the points of the first error subregion, $\mathcal{F}_0$, have a value of q_2 which is between $\vartheta_2(q_1)$ and ϑ_1. For $k > 0$, $\mathcal{F}_k$ comprises two parts, one above and one below the $\vartheta_2(q_1)$ curve. Thus, we can write:

$$\begin{aligned}\varphi_k &= \int_{\alpha_k}^{\alpha_{k+1}} |\vartheta_2(q_1) - \alpha_{N+1-k}| \mathrm{d}q_1 = \\ &= \int_{\alpha_k}^{\vartheta_2(\alpha_{N+1-k})} (\vartheta_2(q_1) - \alpha_{N+1-k}) \mathrm{d}q_1 + \\ &+ \int_{\vartheta_2(\alpha_{N+1-k})}^{\alpha_{k+1}} (\alpha_{N+1-k} - \vartheta_2(q_1)) \mathrm{d}q_1\end{aligned} \tag{4}$$

The symmetry of the curve ϑ_2 implies that $\vartheta_2(\vartheta_2(q)) = q$. By exploiting this property in the relationship $E = \sum_{k=0}^{N} \varphi_k$ one can take the first order derivative with respect to α_k, obtaining:

$$\mathrm{d}E/\mathrm{d}\alpha_k = 2\alpha_{N+1-k} - 4\vartheta_2(\alpha_k) + 2\alpha_{N+2-k} ,$$

where all the resulting terms $\vartheta_2(\vartheta_2(\alpha_k)) - \alpha_k$ are equal to 0. By imposing all derivatives to be equal to 0, (2) is obtained. This is also shown to correspond to a minimum as the second order derivative is $\mathrm{d}^2E/\mathrm{d}\alpha_k^2 = -4(\mathrm{d}\vartheta_2/\mathrm{d}\alpha_k) > 0$.

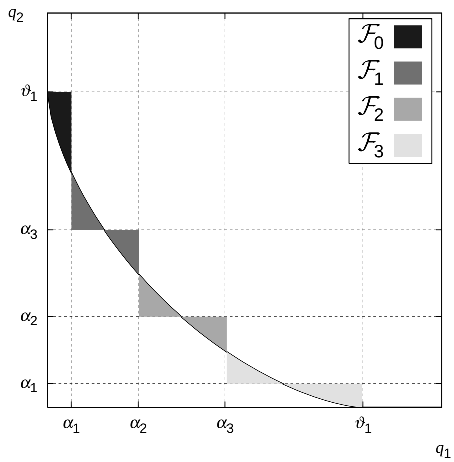

Fig. 7. Thresholds and area error regions of the i-approach for $N = 3$.

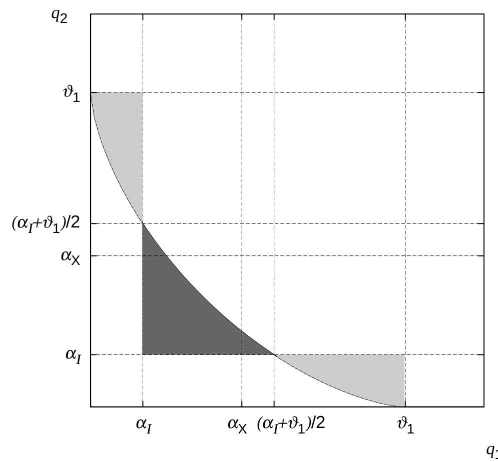

Fig. 8. Single-threshold case of the i-approach and resulting area errors.

APPENDIX B
COMPARISON OF I- AND X- APPROACHES

Given a function $\vartheta_2 : [0,1] \to [0,1]$ which is decreasing and concave, there are two possibilities to approximate the regions identified by the curve in $[0,1]^2$ traced by $(q_1, q_2) : q_2 = \vartheta_2(q_1)$ by means of rectangular regions which are obtained as described in Section III. For the sake of simplicity, we give the proof only for $N = 1$, i.e., a single threshold is added between 0 and ϑ_1. This situation is depicted in Figs. 8 and 9 for the i-approach and the x-approach, respectively. The extension to $N > 1$ is straightforward.

Let us call α_I the optimal threshold in the i-approach. As a consequence of (2), $\vartheta_2(\alpha_I) = \frac{1}{2}(\alpha_I + \vartheta_1)$. For the x-approach instead, the only optimal threshold α_X can be shown to satisfy $\vartheta_2(\alpha_X) = \alpha_X/2$. It is also $\alpha_X > \alpha_I > \alpha_X/2$. As a result, points (α_X, α_X) and (α_I, α_I) are always above and below the curve $\vartheta_2(q_1)$, respectively, as shown in Figs. 8 and 9. The proofs of these statements can be derived following the same approach of Appendix A.

From Fig. 8, observe that the error area term marked with a darker shade *minus* the error areas marked with a lighter shade equals $\int_0^{\vartheta_1} \vartheta_2(q_1)\mathrm{d}q_1 + \alpha_I^2 - 2\alpha_I\vartheta_1$.

Therefore, the area error of the i-approach is

$$\mathcal{E}_I \;=\; \int_0^{\vartheta_1} \vartheta_2(q_1)\mathrm{d}q_1 + \alpha_I^2 + 2\alpha_I\vartheta_1 - 4\int_0^{\alpha_I} \vartheta_2(q_1)\mathrm{d}q_1.$$

Since because of the symmetry of the curve (see Fig. 8)

$$\int_0^{\alpha_I} \vartheta_2(q_1)\mathrm{d}q_1 = \alpha_I(\alpha_I + \vartheta_1)/2 + \int_{\frac{\alpha_I+\vartheta_1}{2}}^{\vartheta_1} \vartheta_2(q_1)\mathrm{d}q_1$$

we have

$$\mathcal{E}_I \;=\; \int_0^{\vartheta_1} \vartheta_2(q_1)\mathrm{d}q_1 - \alpha_I^2 - 4\int_{\frac{\alpha_I+\vartheta_1}{2}}^{\vartheta_1} \vartheta_2(q_1)\mathrm{d}q_1\,.$$

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

$$\mathcal{E}_X \;=\; 2\int_{\alpha_X}^{\vartheta_1} \vartheta_2(q_1)\mathrm{d}q_1 + \frac{\alpha_X^2}{2} - \int_{\alpha_X/2}^{\alpha_X} \vartheta_2(q_1)\mathrm{d}q_1,$$

where the first and the remaining terms comprise the regions with light grey and dark grey shade, respectively, in Fig. 9.

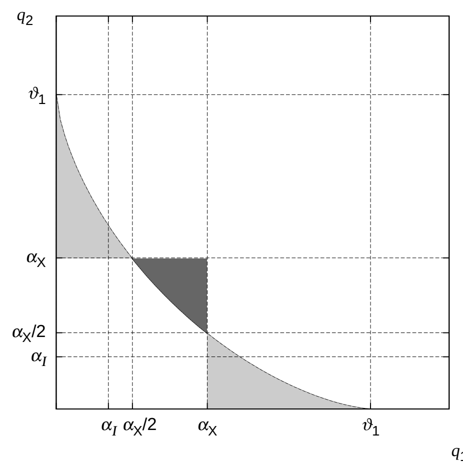

Fig. 9. Single-threshold case of the x-approach and resulting area errors.

We can re-arrange this expression by observing that:

$$\int_0^{\alpha_X/2} \vartheta_2(q_1)\mathrm{d}q_1 = \int_{\alpha_X}^{\vartheta_1} \vartheta_2(q_1)\mathrm{d}q_1 + \alpha_X\frac{\alpha_X}{2}\,,$$

and therefore

$$\mathcal{E}_X \;=\; \int_0^{\vartheta_1} \vartheta_2(q_1)\mathrm{d}q_1 - 2\int_{\alpha_X/2}^{\alpha_X} \vartheta_2(q_1)\mathrm{d}q_1\,.$$

Thus, to end the proof we need to show that

$$\alpha_I^2 + 4\int_{(\alpha_I+\vartheta_1)/2}^{\vartheta_1} \vartheta_2(q_1)\mathrm{d}q_1 \geq 2\int_{\alpha_X/2}^{\alpha_X} \vartheta_2(q_1)\mathrm{d}q_1. \qquad (5)$$

Because of concavity, $\vartheta_2'(\frac{\alpha_I+\vartheta_1}{2}) \geq -1$, so that

$$\int_{(\alpha_I+\vartheta_1)/2}^{\vartheta_1} \vartheta_2(q_1)\mathrm{d}q_1 \geq \alpha_I^2/2.$$

The concavity of the curve also implies that the region below ϑ_2 between $\alpha_X/2$ and α_X is all contained within the trapezoid with vertices $(\alpha_X/2, 0)$, $(\alpha_X, 0)$, $(\alpha_X, \alpha_X/2)$, and $(\alpha_X/2, \alpha_X)$, whose area is $\frac{3}{8}\alpha_X^2$. Therefore, (5) is proved since its left-hand member is not less than $3\alpha_I^2 \geq \frac{3}{4}\alpha_X^2$ which is in turn not less than the right-hand member.

REFERENCES

[1] 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. Inf. Theory*, vol. 47, no. 6, pp. 2275–2299, Sep. 2001.

[2] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error–correcting coding and decoding," in *Proc. IEEE ICC*, vol. 2, Geneva, Switzerland, Nov. 1993, pp. 1064–1070.

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

[4] 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.

[5] 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.

[6] W. Turin, *Performance Analysis of Digital Transmission Systems*. New York: Computer Science Press., 1990.

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

[8] 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, Jun. 2006.

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

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

[11] T. M. Cover and J. A. Thomas, *Elements of Information theory*. New York: John Wiley & Sons, INC., 1991.