This document is a preprint of: W. Rave, "Quantization of Log-Likelihood Ratios to Maximize Mutual Information," in IEEE Signal Processing Letters (SPLet), vol. 16, no. 4, Apr 2009. DOI:10.1109/LSP.2009.2014094

Quantization of Log-Likelihood Ratios to Maximize Mutual Information

Wolfgang Rave
Vodafone Chair Mobile Communications Systems, Technische Universität Dresden
D-01062 Dresden, Germany Email: {rave}@ifn.et.tu-dresden.de

***Abstract*—We propose a quantization scheme for log-likelihood ratios which optimizes the trade-off between rate and accuracy in the sense of rate distortion theory: as distortion measure we use mutual information to determine quantization and decision levels maximizing mutual information for a given rate over a Gaussian channel. This approach is slightly superior to the previously proposed idea of applying the *Lloyd-Max* algorithm to the 'soft bit' density associated to the L-values. A further data rate reduction can be achieved with *entropy coding*, because the optimum quantization levels based on mutual information are used with pronounced unequal probabilities.**

***Index Terms*—Mutual information, soft bits, quantization, entropy coding, iterative decoding**

I. Introduction

In many signal processing applications concatenated modules exchange probabilities to perform certain tasks in a given signal processing chain or even do this iteratively to process the data. For digital processing the signals are typically represented by bits in binary form and so-called log-likelihood ratios (abbreviated as LLRs or L-values in the following) occur and have to be communicated with finite precision. The necessary quantization step inevitably causes quantization noise equivalent to information loss in the system.

In [1] we presented a quantization scheme of log-likelihood ratios $L(X)$ using 'soft bits' $\Lambda(X)$ related to coded or information bits X and their associated L-values through $\Lambda(X) = E\{X\} = \tanh(L(X)/2)$ [2]. The underlying idea is the intuitive consideration, that a (typically Gaussian) density that extends to $\pm\infty$ is more difficult to quantize than some transformed variable that shows saturation with increasing reliability (magnitude) of the L-value. Based on a closed form expression for the transformed soft bit density that we had derived, $Lloyd$'s optimum algorithm for scalar quantization [3], [4] could be conveniently applied to derive *decision* and *reconstruction* (quantization) levels d_i and r_i, respectively. We also showed that the loss of mutual information due to quantization was clearly smaller with soft bit based quantization than with 'direct' quantization of the L-value density.

One might ask, whether the mean square error (or some more general norm) is really the best optimization criterion in such a case. Recognizing that the loss in mutual information (w.r.t. to unquantized L-values) can be taken as the cost function to determine the quantizer levels, it is possible to replace the MSE minimization employed by *Lloyd*'s optimum scalar quantization to the soft bit density by a direct minimization of the mutual information loss. This leads to our new quantization scheme: a maximization of mutual information of the quantized L-value density w.r.t. the d_i and r_i.

In the following we carry out this idea. Taking the optimum scalar quantization of L-value and soft bit densities as a reference (Section II) the mutual information between the binary variable X and its log-likelihood ratio L is maximized in Section III. The possibility to further reduce the required data rate using entropy coding is demonstrated in Section IV. A performance example is given by decoding a turbo code with quantized L-values in Section V.

II. Scalar Quantization of L-values and Associated Soft-Bit Densities

Assume that after processing a received signal, bit probabilites (e.g. a-posteriori probabilities at the decoder output) represented as log-likelihood ratios have to be communicated to another module or location. Let us also assume a Gaussian channel[1] $y = x + n$ with noise variance σ_n^2 for which the L-values are related to the channel output as $l = (2/\sigma_n^2)y$. The conditional density[2] $p_{L|X}(l|x)$, conditioned on one of the channel inputs $x = +1$ or $x = -1$, is also Gaussian [5], [6] with variance $\sigma_L^2 = 4/\sigma_n^2$. Furthermore it depends only on the single parameter σ_L, because mean and variance are related by $\mu_L = \sigma_L^2/2$. Thus the unconditioned L-value density for equally likely input values $x = \pm 1$ follows a bimodal Gaussian distribution[3]:

$$\begin{aligned} p_L(l) = \quad & \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_L}\left[\exp\left(-\frac{(l-\frac{\sigma_L^2}{2})^2}{2\sigma_L^2}\right)+\exp\left(-\frac{(l+\frac{\sigma_L^2}{2})^2}{2\sigma_L^2}\right)\right] \\ = \quad & \frac{1}{2}\left[\mathcal{N}(\tfrac{\sigma_L^2}{2},\sigma_L^2)+\mathcal{N}(-\tfrac{\sigma_L^2}{2},\sigma_L^2)\right] \end{aligned} \quad (1)$$

Intuitively one expects, that the precision with which a certain L-value has to be represented by the quantizer varies with its magnitude: L-values of large magnitude need *not* be represented as precisely as small ones, while the transition range from low to high reliability might benefit from increased reconstruction level density.

Therefore the description of L-values in terms of '*soft bits*' appears appropriate: while the Gaussian L-value density extends to $\pm\infty$, soft bits $\lambda(l)$ related to L-values by

[1] This includes AWGN and flat Rayleigh channels, but also holds approximately due to the central limit theorem for large block size in general.

[2] We denote random variables such as X, $L(X)$ and $\Lambda(X)$ with capital letters and their realizations x, l and λ in lower case.

[3] The variance of this density is $\sigma_L^2 + \sigma_L^4/4$. We will use, however, σ_L^2 the variance of the *conditional* density as parameter throughout.

the hyperbolic tangent only cover the range $[-1, 1]$. Scalar quantization according to *Lloyd* [3] and *Max* [4], optimum in the sense of minimizing the mean square error (MSE) between the quantized and non-quantized densities, therefore should rather be applied to the soft bit density, where levels will not be wasted at a wide range of large magnitudes to keep the quadratic error small, while the associated change in reliability is already quite small.

Using the standard normal distribution $\mathbb{G}(\eta) = (1/\sqrt{2\pi})\int_{-\infty}^{\eta} \mathrm{e}^{-\xi^2/2}d\xi$ we obtained the probability $\Pr(L(X) \leq \eta)$ which can be converted with the inverse soft bit $L(X) = 2\tanh^{-1}(\Lambda(X))$ to the equivalent probability $\Pr(\tanh^{-1}(\Lambda(X)) \leq \eta)$ expressed in terms of soft bits. Taking the derivative of this distribution function we derived the soft bit density $p_\Lambda(\lambda)$ in closed form as [1]:

$$p_\Lambda(\lambda) = \frac{1}{(1-\lambda^2)\sqrt{2\pi}\sigma_L}\left[\exp\left(-\frac{(4\tanh^{-1}\lambda-\sigma_L^2)^2}{8\sigma_L^2}\right)\cdots + \exp\left(-\frac{(4\tanh^{-1}\lambda+\sigma_L^2)^2}{8\sigma_L^2}\right)\right]. \quad (2)$$

Examples of the involved densites defined by Eqs. (1), (2) are illustrated in [1], where also plots of the optimized reconstruction and decison levels r_i and d_i, determined with optimum scalar quantization for different numbers $b = \mathrm{ld}\,R$ of quantizer bits[4], as a function of the variance σ_L^2 of the (conditioned) L-values can be found.

III. Minimization of the Mutual Information Loss due to Quantization

Given the sets r_i and d_i the mutual information $I(X; L)$ between the binary variable $X \in \{-1, +1\}$ and its quantized (discrete) and non-quantized continuous L-value densities can be calculated. For a continuous density using the symmetry and consistency [5] of L-values we have

$$\begin{aligned} I_\mathrm{c}(X; L) &= \sum_{x=\pm 1}\int_{-\infty}^{\infty}\frac{p(l|x)}{2}\mathrm{ld}\frac{2p(l|x)}{p(l|x=1)+p(l|x=-1)}dl \\ &= \frac{1}{2}\sum_{x=\pm 1}\int_{-\infty}^{\infty} p(l|x)\,[1-\mathrm{ld}(1+e^{-lx})]dl\,. \quad (3)\end{aligned}$$

For a Gaussian channel the integration in the discrete (quantized) case is easily carried out. Expanding the sum over the two possible bit levels[5] of the information bit $X \in \{\pm 1\}$ the mutual information between X and its quantized log-likelihood ratio $L(X)$ can be written as[6]

$$\begin{aligned} I_\mathrm{q}(X; L) &= \frac{1}{4}\sum_{i=1}^{R}[1-\mathrm{ld}(1+e^{-r_i})]\,\mathrm{erf}\left(\frac{l-\mu_L}{\sqrt{2}\sigma_L}\right)\Bigg|_{d_i}^{d_{i+1}}\cdots \\ &\quad + [1-\mathrm{ld}(1+e^{r_i})]\,\mathrm{erf}\left(\frac{l+\mu_L}{\sqrt{2}\sigma_L}\right)\Bigg|_{d_i}^{d_{i+1}}. \quad (4)\end{aligned}$$

[4]Logarithms w.r.t. base 2 and natural logarithms are denoted as $\mathrm{ld}(\cdot)$ and $\ln(\cdot)$ throughout the paper.

[5]Due to symmetry the sum over the two bit levels actually can be skipped; it is only relevant when taking derivatives.

[6]The standard normal distribution $\mathbb{G}(\eta)$ could here be used again. We used the error function for practical evaluation instead, because it is commonly available in math packages.

The difference between the two mutual informations represents the *information loss* due to quantization: $\Delta I = I_\mathrm{q} - I_\mathrm{c}$. To minimize this loss we use the derivatives of I_q. W.r.t. the reconstruction levels r_i $(1 \leq i \leq R)$ they are given by

$$\begin{aligned} \frac{\partial I_\mathrm{q}}{\partial r_i} &= \frac{1}{4\ln 2}\frac{e^{-r_i}}{1+e^{-r_i}}\,\mathrm{erf}\left(\frac{l-\mu_L}{\sqrt{2}\sigma_L}\right)\Bigg|_{d_i}^{d_{i+1}}\cdots \\ &\quad -\frac{1}{4\ln 2}\frac{e^{r_i}}{1+e^{r_i}}\,\mathrm{erf}\left(\frac{l+\mu_L}{\sqrt{2}\sigma_L}\right)\Bigg|_{d_i}^{d_{i+1}}. \quad (5)\end{aligned}$$

The derivatives w.r.t. the decision levels d_i $(1 \leq i \leq R+1)$ are given by

$$\begin{aligned} \frac{\partial I_\mathrm{q}}{\partial d_i} &= \frac{1}{2\sqrt{2\pi}\sigma_L}\mathrm{ld}\frac{1+e^{-r_i}}{1+e^{-r_{i-1}}}\exp\left(-\frac{(d_i-\mu_L)^2}{2\sigma_L^2}\right)\cdots \\ &\quad + \frac{1}{2\sqrt{2\pi}\sigma_L}\mathrm{ld}\frac{1+e^{r_i}}{1+e^{r_{i-1}}}\exp\left(-\frac{(d_i+\mu_L)^2}{2\sigma_L^2}\right). \quad (6)\end{aligned}$$

Optimum levels in terms of mutual information loss were determined by maximizing the mutual information I_q of the quantized L-value density by a steepest ascent technique (we alternatingly optimized the two level types in the spirit of *Lloyd*'s algorithm for several iterations).

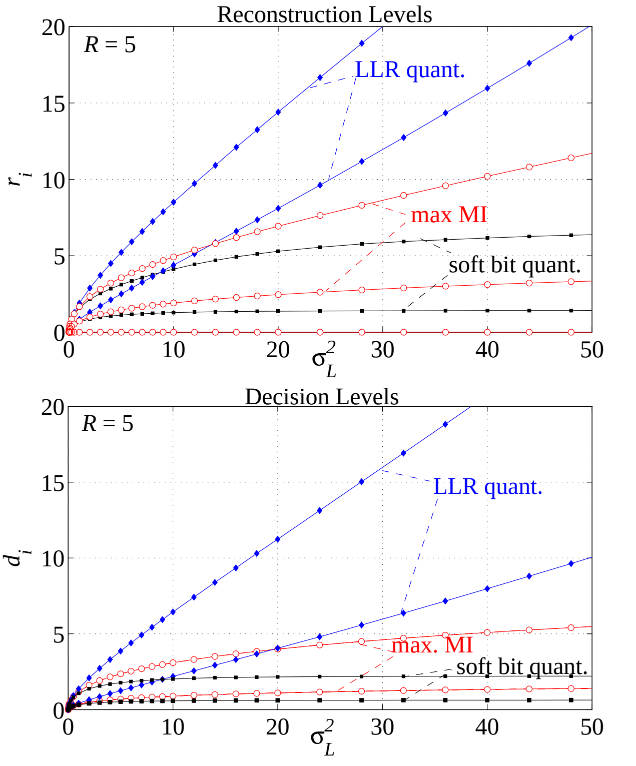

Fig. 1. Comparison of optimized reconstruction and decision levels for $R = 5$ levels: *Lloyd*'s algorithm applied to LLR- and soft bit densities vs. maximisation of mutual information.

Example results are presented for $R = 5$ reconstruction levels (4 finite decision levels; the decision levels d_1 and d_{R+1} have the values $\pm\infty$ independent of the method of quantization) in Fig. 1. Due to symmetry two level pairs occur for both reconstruction and decision levels which only differ in sign while the 5^{th} reconstruction level is exactly zero independent of the L-value variance. Therefore only positive levels are shown as a function of σ_L^2 comparing LLR and soft bit quantization with the scheme maximizing mutual information.

We note that the levels due to soft bit quantization show the expected saturation with increasing L-value magnitude (increasing variance σ_L^2). This effect is less pronounced after maximizing the mutual information I_q. Apart from optimizing the levels directly w.r.t. the nonlinear expression for the mutual information, a new degree of freedom that can be exploited with the latter scheme is the independent optimization of the decision levels which do not have to be the mean of the neighboring reconstruction levels (corresponding to 1D-Voronoi regions, when viewed as a special case of vector quantization). This allows to reduce the loss in mutual information as shown in Fig. 2.

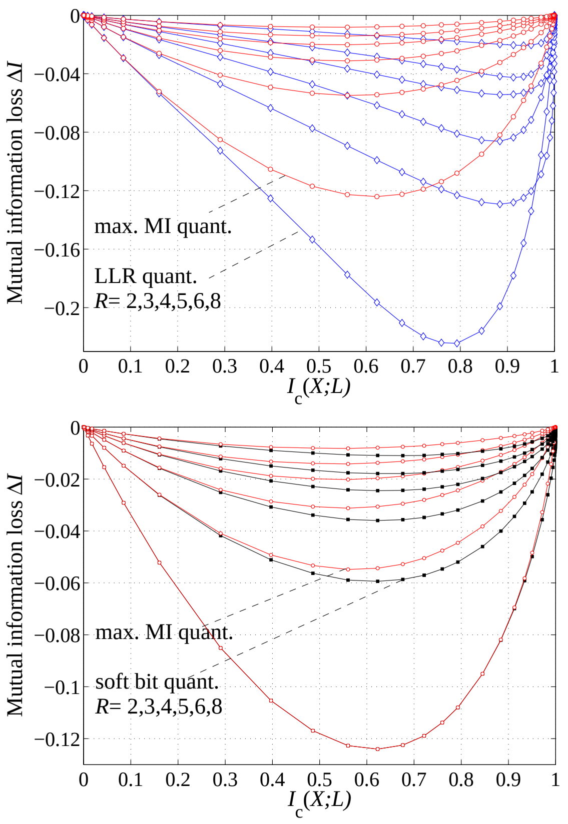

Fig. 2. Comparison of the mutual information loss ΔI between quantized and not quantized L-values as a function of the average mutual information of the L-values: LLR vs. mutual information optmization and soft bit vs. mutual information optimization for $R = 2 - 8$ reconstruction levels.

We note a relatively large reduction of the mutual information loss (decreasing with increasing number of levels), when the LLR-based quantization is replaced by the maximization of the mutual information. This is shown for $R = 2, 3, 4, 5, 6, 8$ reconstruction levels. The comparison between the soft bit quantisation and the mutual information maximization shows that with soft bits a large portion of the loss can already be avoided (note the different scale of the ordinates in Fig. 2), although still up to 1/4 of this loss can be saved with the separate optimization of the decision levels. Here the case

fitting parameter	a	b	c
$R = 2, r_1$	0.4473	0.2806	0.1075
$R = 3, r_1$	0.6865	0.4861	0.0955
$R = 3, d_1$	0.218	0.4082	-0.01687
$R = 4, r_1$	0.1305	0.3251	-0.01053
$R = 4, r_2$	0.8214	0.6343	0.08861
$R = 4, d_1$	0.33	0.6509	-0.02219
$R = 5, r_1$	0.2361	0.5248	-0.01938
$R = 5, r_2$	0.9179	0.7457	0.08219
$R = 5, d_1$	0.1605	0.2223	-0.01146
$R = 5, d_2$	0.3924	0.8367	-0.02881
$R = 6, r_1$	0.1299	0.1831	-0.009314
$R = 6, r_2$	0.3312	0.6499	-0.02541
$R = 6, r_3$	1.002	0.8184	0.07718
$R = 6, d_1$	0.2813	0.3667	-0.01879
$R = 6, d_2$	0.4546	0.9589	-0.03378
$R = 8, r_1$	0.1358	0.09532	-0.006295
$R = 8, r_2$	0.3794	0.3377	-0.01972
$R = 8, r_3$	0.5314	0.7521	-0.03236
$R = 8, r_4$	1.179	0.8661	0.07178
$R = 8, d_1$	0.2737	0.199	-0.01278
$R = 8, d_2$	0.4993	0.4999	-0.02727
$R = 8, d_3$	0.611	1.063	-0.0398

TABLE I
FITTING PARAMETERS FOR $R = 2, 3, 4, 5, 6, 8$ RECONSTRUCTION LEVELS.

$R = 2$ is special, because this degree of freedom is not yet available, so that both schemes are equivalent.

For hardware implementation a look-up table with interpolation could be used. More practical, however, is to use curve fitting based on an appropriate approximating function. In this case an expansion in terms of σ_L works well. Fits with $f(\sigma_L) = a\sqrt{\sigma_L} + b\sigma_L + c\sigma_L^2)$ provided mean square errors in the range $0 \le \sigma_L^2 \le 100$ of the order 10^{-3}. Table I collects the coefficients for $R = 2 - 8$ reconstruction levels.

IV. Additional Rate Reduction with Entropy Coding

Another observation concerns the probabilities P_i with which different quantization levels are used: while quantization of the LLR density with the *Lloyd-Max* algorithm leads to more or less uniformly used levels, L-value quantization based on mutual information or soft bits causes significant differences in probability which depends on σ_L^2.

This can be exploited with a well-known technique referred to as *entropy coding* [7], [8] that assigns quantization labels of different length using a prefix-free code either to the individual levels or to groups of them, if several L-values are represented jointly. A lower bound for the required average label length is given by the entropy of the quantization levels $H_R(L_q) = -\sum_{i=1}^{R} P_i \log P_i + (1 - P_i)\log(1 - P_i)$.

We compare this bound for LLR quantization based on mutual information and the *Lloyd-Max* algorithm applied to the L-value density in Fig. 3 again as a function of σ_L^2. In addition we computed the practically achievable rate reduction using Huffman coding in 3 dimensions. This corresponds to the staircase functions lying above the bounding entropy curves.

This document is a preprint of: W. Rave, “Quantization of Log-Likelihood Ratios to Maximize Mutual Information,” in IEEE Signal Processing Letters (SPLet), vol. 16, no. 4, Apr 2009. DOI:10.1109/LSP.2009.2014094

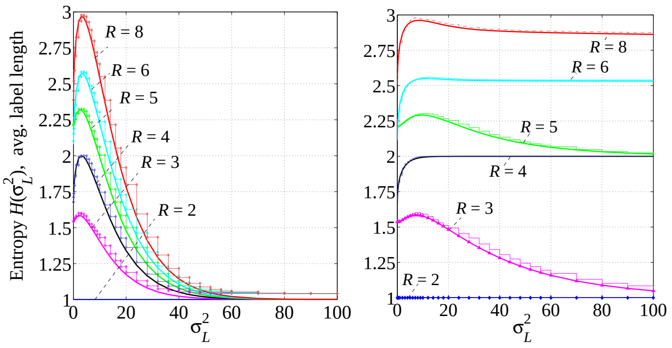

Fig. 3. Entropy comparison of L-values quantized based on mutual information (left) or the *Lloyd-Max* algorithm (right) as a function of σ_L^2 and the average label length achievable with 3D Huffman coding.

V. Application of Quantized L-Values

As a test case for the proposed LLR quantization scheme we studied decoding of a parallel concatenated convolutional code (PCCC). The extrinsic information obtained from the decoder output with the BCJR algorithm (applying the max. Log-MAP approximation) was exchanged either after direct LLR quantization, soft bit quantization or quantization based on maximizing the mutual information for $R = 2, 3, 4, 8$ levels by the two companion decoders. As a reference the unquantized representation of the a-priori information was taken. We used a simple recursive systematic PCCC with generators $G = [1, 5/7]_8$ block length $N = 2000$ with a random interleaver punctured to obtain rate 1/2. Results comparing the three quantization schemes are presented in Fig. 4.

The variance of the LLRs was adaptively determined at the MAP decoder output from the observed variance of $L(X)$, the LLR of the code bits. The fitting functions (the coefficients from Table I) were used to obtain the new quantization levels in each decoder iteration. For variances larger than $\sigma_L^2 = 100$ the levels for that value were taken and the current L-value distribution was scaled accordingly to maintain the information about the relative magnitude among the L-values (one could stop iterating here, too). The findings here can be compared with the observations in [9], where quantization of all involved quantities in the Log-MAP algorithm was studied. The results for LLR quantization in the upper plot are markedly worse than the other two schemes, because quantization levels are wasted at large L-value magnitude. In the lower plot the soft bit quantization (dashed lines) is slightly worse than the mutual information based technique, consistent with Fig. 2.

VI. Concluding Remarks

Distinguishing three quantization schemes, we showed that a significant gain is possible, if LLR values are not quantized ’directly’ with the *Lloyd-Max* algorithm. Rather soft bits or even slightly better mutual information provide better cost functions. We quantified this with the loss in mutual information, demonstrating the improvement with a turbo decoding example. Although the proposed scheme strictly applies only to Gaussian channels, it should also be of interest

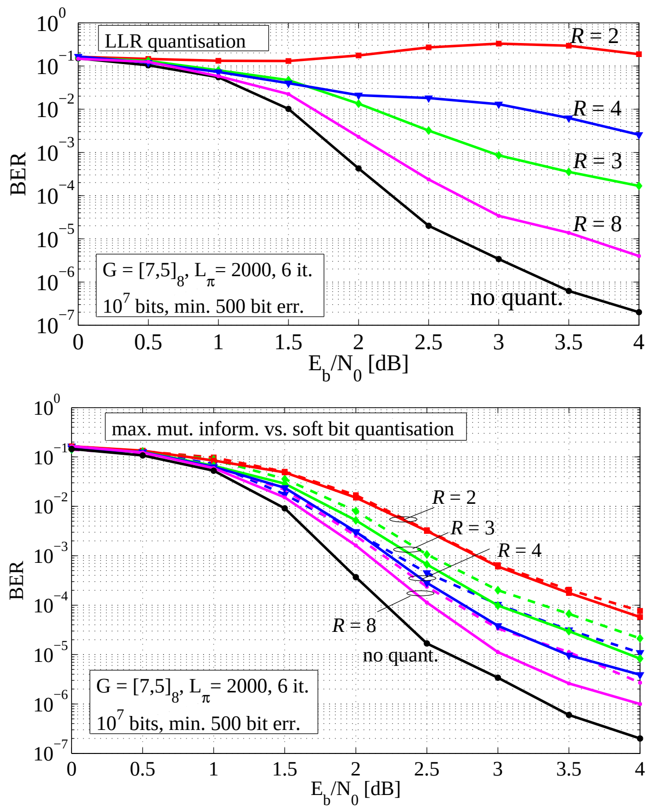

Fig. 4. Log-MAP decoding of a PCC code ($G = [1, 5/7]_8$, blocklength $N = 2000$, AWGN channel) with quantized extrinsic information.

for coded transmission over fading channels, if a Gaussian approximation for the L-value density at the decoder output is applicable. An extension from scalar to vector quantization is feasible by generalizing Eqs. (3) and (4) to vectors. The interesting question, how much can be gained by the additional degrees of freedom from more dimensions remains as an open problem for future work.

References

[1] S. Khattak, W. Rave and G. Fettweis, *“Distributed Iterative Multiuser Detection through Base Station Cooperation,”*, EURASIP J. on Wireless Comm. and Networking Volume 2008, Article ID 390489

[2] J. Hagenauer, E. Offer, L. Papke, *“Iterative decoding of binary block and convolutional codes”*, IEEE Trans. Inf. Theory, vol.42, pp. 429-445, 1996

[3] S. Lloyd, *”Least squares quantization in PCM,”* (unpubl. Bell Lab. Techn. Note, 1957) IEEE Trans. Inf. Theory, vol. IT-28, pp. 129 - 137, 1982

[4] J. Max, *”Quantizing for minimum distortion”*, IEEE Trans. Inf. Theory, vol. 6, pp. 7-12, 1960

[5] J. Hagenauer, *“The Exit Chart (Invited Paper)”*, Proc. 12th Europ. Signal Proc. Conf. (EUSIPCO), pp. 1541-1548, 2004

[6] S. ten Brink, *”Convergence behavior of iteratively decoded parallel concatenated codes,”* , IEEE Trans. Comm., vol. 49., pp. 1727-37 , 2001

[7] A. Gersho and R. M. Gray, *”Vector Quantization and Signal Compression,”* Boston, MA: Kluwer, 1992

[8] Thomas M. Cover and Joy A. Thomas, *”Elements of Information Theory”* 2^{nd} Ed., John Wiley & Sons, 2006

[9] P. Robertson, P. Hoeher, E. Villebrun, *“Optimal and sub-optimal maximum a posteriori algorithms suitable for turbo decoding”*, Europ. Trans. on Telecomm. , vol. 8, pp. 119-125, 1997