# On the Kullback-Leibler Distance and the Mean Square Distortion of Mismatched Distributed Quantizers

Gerhard Maierbacher João Barros

Department of Computer Science and
Laboratory of Artificial Intelligence and Computer Science
Universidade do Porto
R. Campo Alegre 823, 4150-180 Porto, Portugal
Email: {gerhard, barros}@ncc.up.pt

***Abstract*— We formulate a problem considering the implications of conditional mean estimation based on unknown or approximated source statistics with respect to the mean squared error of the estimates. Since we are specially interested in a solution for a multiterminal quantization problem where the statistical properties of the data to be estimated might play a key role towards finding a mathematical solution, we formulate the problem based on this concrete scenario without compromising a more general treatment of the problem. Specifically, we are interested in establishing a link between the mean squared error and the mismatch of used probability density function measured by the Kullback-Leibler distance.**

***Index Terms*— Multiterminal quantization, distributed source coding, conditional mean estimation, PDF mismatch**

## I. Introduction

In this paper we formulate a problem considering the implications of conditional mean estimation (CME) based on unknown or approximated source statistics onto the mean squared error (MSE) of the estimates. Specifically, we address a distributed source coding scenario where correlated data from a large number of independently operating encoders has to be decoded in a centralized fashion (based on CME), such that the MSE distortion after joint decoding is minimized. The considered system model is depicted in Figure 1. Assuming that each encoder within the scenario uses a simple scalar quantizer (Block Q in Figure 1) to discretize its measurements and a simple mapping table (block BM in Figure 1) to determine the binary codewords to be transmitted, we focus on a joint decoding concept able to exploit the correlations between the encoder inputs to minimize the end-to-end distortion within the system. Since in reality the actual joint probability density function (PDF) of the encoder inputs is not available, the decoder has to rely on an approximated PDF, which generally differs from the original PDF. The difference between both PDF's is also denoted as *mismatch* and can be quantified using the Kullback-Leibler distance (KLD), see [1]. The results in [5] indicate that there exists a useful connetion between the MSE distortion (the original figure of merit) and the mismatch between the actual and the approximated PDF used at the decoder. Since we have not been able to find an exact mathematical proof for this so far, the main purpose of this work is to provide a rigorous formulation of the problem.

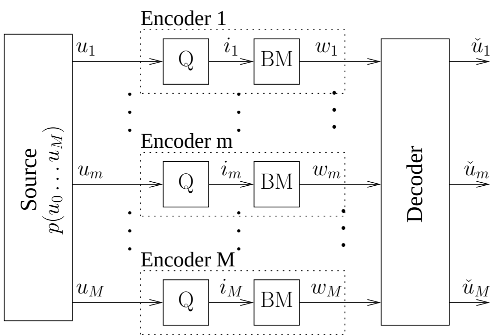


Fig. 1. Considered System Model.

The rest of this paper is organized as follows: In Section II we present the source model considered in our scenario and provide some explanations to the encoding procedure, which might provide some important clues for a mathematical solution of the problem. In Section III we present different decoding concepts (based on CME), which are needed for the problem formulation. The Kullback-Leibler distance is addressed in Section IV and the problem itself will be stated in Section V.

## II. Multiterminal Quantization

We start by introducing our notation: Random variables are always denoted by capital letters $U$, where its realization is denoted by the corresponding lowercase letter $u$. Vectors of random variables are denoted by bold capital letters and always assumed to be column vectors $\mathbf{U} = (U_1, U_2, \cdots, U_M)^T$, whereas its realization is denoted by the corresponding bold lowercase letter $\mathbf{u} = (u_1, u_2, \cdots, u_M)^T$. Sets are denoted by capital caligraphic letters, e.g. $\mathcal{M}$, where $|\mathcal{M}|$ denotes the set's cardinality. The covariance is defined by $Cov\{\mathbf{a}, \mathbf{b}\} \triangleq E\{\mathbf{a}\mathbf{b}^T\} - E\{\mathbf{a}\}E\{\mathbf{b}\}^T$, where $E\{\cdot\}$ is the expectation operator. An $M$-dimensional random variable with realizations $\mathbf{u} = (u_1\, u_2 \dots u_M)^T$, $u_m \in \mathbb{R}$, is Gaussian distributed with mean $\boldsymbol{\mu} = E\{\mathbf{u}\}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}\{\mathbf{u}, \mathbf{u}\}$, when its PDF $p(\mathbf{u})$ is given by.

$$p(\mathbf{u}) = \exp(-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu}))/(2\pi|\boldsymbol{\Sigma}|)^{1/2}. \qquad (1)$$

Such a PDF is simply denoted as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

### A. Source Model

Each encoder $m$, with $m = 1, 2, \cdots, M$, observes a continuous-valued source sample $u_m(t)$ at time $t$. In this work, we only consider the spatial correlation of measurements and not their temporal dependence so that the time index $t$ is dropped and only one time step is considered. The sample vector $\mathbf{u} = (u_1, u_2, \cdots, u_M)^T$ at any given time $t$ is assumed to be one realization of a $M$-dimensional Gaussian random variable, whose PDF $p(\mathbf{u})$ is given by $\mathcal{N}(\mu, \Sigma)$ with

$$\Sigma = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,M} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{M,1} & \rho_{M,2} & \cdots & 1 \end{bmatrix}$$

and $\mu = \mathbf{0}_M$, where $\mathbf{0}_M$ denotes the length-$M$ all-zero column vector.

### B. Quantization

Since the observation $u_m$ at sensor $m$, with $m = 1, \cdots, M$, is drawn from a continuous-valued distribution, the first step in digital data processing is to convert these observations into a discrete representation. This happens at the quantizer, see block $Q$ in Figure 1.

In this work we will restrict our attention to scalar quantizers. Such quantizers take a single value $u_m$ and map this value onto a discrete index $i_m$, which can take values within a discrete alphabet $\mathcal{I}_m = \{0, 1, \cdots, |\mathcal{I}_m| - 1\}$. An exemplary quantization characteristic is shown in Figure 2. During quantization an input value $u_m$ is mapped onto the index $i_m = l$ if it falls within the interval $\mathcal{B}_m(l)$ between the decision levels $b_m(l)$ and $b_m(l+1)$, such that $b_m(l) < u_m \leq b_m(l+1)$. The obtained quantization index $i_m$ is then associated with

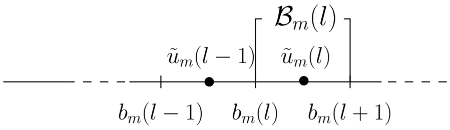


Fig. 2. Quantization of $u_m$

the reconstruction level $\tilde{u}_m(l)$, representing all source samples $u_m$ falling into the interval $\mathcal{B}_m(l)$. This is shown in Figure 2. The set of reconstruction levels $\mathcal{U}_m$ consists of the elements $\mathcal{U}_m = \{\tilde{u}_m(0), \cdots, \tilde{u}_m(|\mathcal{I}_m| - 1)\}$. The quantization function is denoted by $q_m(u_m)$, where

$$i_m = q_m(u_m). \tag{2}$$

The quantizers at the encoders are designed to minimize the mean squared error (MSE) $\delta_m$ within the observations, where

$$\begin{aligned} \delta_m &= E\{||U_m - \tilde{U}_m(\mathbf{I}_m)||^2\} \\ &= \int_{-\infty}^{\infty} (u_m - \tilde{u}_m(i_m))^2 \cdot p(u_m)\ du_m. \end{aligned} \tag{3}$$

In general, there are two different types of (scalar) quantizers - *Uniform quantizers* and *non-uniform quantizers*. An excellent description of quantizers and the differences between uniform and non-uniform quantization can be found in standard literature [2] or in the paper by Gray and Neuhoff [4].

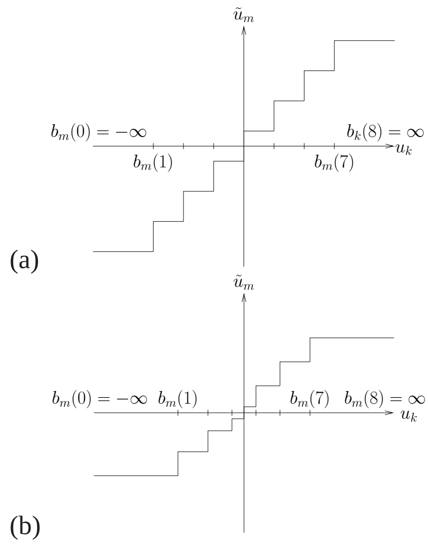


Fig. 3. Quantizer characteristics: (a) uniform, (b) Lloyd-Max

In the following we provide a brief overview for both types of quantizers and provide some simplifying approximations, which might be useful for further considerations.

*1) Uniform Quantization:* Uniform quantizers are characterized by the fact that (a) the quantization intervals $\mathcal{B}_m(i_m)$ are equally broad, i.e. $b_m(l+1) - b_m(l) = \Delta$ for all $l \in \mathcal{I}_m$, and (b) the reconstruction levels $\tilde{u}_m(i_m)$ are placed exactely in the middle between the quantization boundaries, i.e. $\tilde{u}_m(i_m = l) = (b_m(l+1) - b_m(l))/2$. An example for an uniform quantizer is provided in Figure 3 (a).

By applying some simplifications [4], the MSE distortion of uniform quantizers with optimally placed quantization intervals for a single source can be approximated by

$$\delta_m^u \cong \frac{\Delta^2}{12}. \tag{4}$$

*2) Non-uniform Quantization:* Non-uniform quantizers do not oblige the strict rules concerning the choice of the quantization areas $\mathcal{B}_m(i_m)$ and/or the choice of the reconstruction values $\tilde{u}_m(i_m)$, which opens additional degrees of freedom to further optimize the MSE distortion. An important optimization algorithm, which iteratively modifies the reconstruction levels $\tilde{u}_m(i_m)$ and the quantization boundaries $b_m(i_m)$ to construct a (locally optimal) quantizer minimizing the MSE for a given Rate $R_m = \log_2(|\mathcal{I}_m|)$, was proposed by Lloyd and Max [2]. An example for an non-uniform Lloyd-Max quantizer is provided in Figure 3 (b).

The MSE distortion of quantizers constructed with the Lloyd-Max algorithm can be approximated [4] by

$$\delta_m^n \cong \frac{1}{12} \left( \int_{-\infty}^{\infty} (p(u_m))^{\frac{1}{3}}\ du_m \right)^3\ 2^{-2R_m}, \tag{5}$$

which can be simplified for Gaussian sources with variance $\sigma$ to

$$\delta_m^n \cong \frac{1}{12} 6\pi\sqrt{3}\sigma^2 2^{-2R_m}. \tag{6}$$

### C. Codeword Mapping

At the mapping-stage (block BM in Figure 1) the quantization indices $i_m$ are mapped onto codeword indices $w_m$. The codeword indices $w_m$ are drawn from a finite-valued alphabet $\mathcal{W}_m = \{0, \cdots, |\mathcal{W}_m| - 1\}$. The codeword-indices of all sensors are denoted by the vector $\mathbf{w} = (w_1, \cdots, w_M)^T$. In this work we focus on one-to-one mappings from quantization indices $i_m$ to codewords $w_m$, although more generally we might choose $|\mathcal{W}_m| \leq |\mathcal{I}_m|$ to achieve distributed data compression. The codeword mappings are formally represented by a mapping function $f_m(\cdot)$, such that $w_m = f_m(i_m)$ for $m = 1, 2, \cdots, M$.

## III. Decoding based on Conditional Mean Estimation

### A. Optimal Decoder

The decoder uses the codeword vector $\mathbf{w} = (w_1, w_2, \cdots, w_M)^T$ and the available knowledge of the source statistics $p(\mathbf{u})$ to produce best possible estimates $\check{u}_m(\mathbf{w})$ of the samples $u_m$. Assuming, that the mean square error (MSE) distortion

$$\check{d} = E\{||\mathbf{U} - \check{\mathbf{U}}(\mathbf{W})||^2\} \tag{7}$$

is to be minimized at the decoder, conditional mean estimation (CME) [3] should be applied and the optimal estimates $\check{u}_m(\mathbf{w})$ based on the codeword vector $\mathbf{w}$ received can be determined by

$$\check{u}_m(\mathbf{w}) = E\{U_m|\mathbf{w}\} = \int_{-\infty}^{\infty} u_m \cdot p(u_m|\mathbf{w}) \; du_m, \tag{8}$$

which generally requires the knowledge of the PDF's $p(u_m|\mathbf{w})$ for all sources $m = 1, 2, \cdots, M$. Assuming that the decoder has unlimited complexity and knowledge about the quantization functions $q_m(\cdot)$ as well as the mapping functions $f_m(\cdot)$, the PDF's $p(u_m|\mathbf{w})$ can be obtained from the PDF $p(\mathbf{u})$ by integrating along the corresponding quantization intervals. Therefore, the PDF $p(\mathbf{u})$ can be seen as the basis for the conditional mean estimation performed.

### B. Mismatched Optimal Decoder

For many reasons the PDF $p(\mathbf{u})$ might not be available at the decoder, such that the decoder has to perform a conditional mean estimation based on a mismatched PDF $\hat{p}(\mathbf{u})$, which differs from $p(\mathbf{u})$. The estimates resulting from the mismatched decoding are denoted by $\hat{u}_m(\mathbf{w})$ and the MSE distortion for mismatched decoding is given by

$$\hat{d} = E\{||\mathbf{U} - \hat{\mathbf{U}}(\mathbf{W})||^2\}. \tag{9}$$

Similarly to (8) the estimates $\hat{u}_m(\mathbf{w})$ can be calculated by

$$\hat{u}_m(\mathbf{w}) = \int_{-\infty}^{\infty} u_m \cdot \hat{p}(u_m|\mathbf{w}) \; du_m. \tag{10}$$

### C. Discrete Decoder

Since the optimal decoder is not feasible (mainly because of the integration operations required), we relax the fidelity criterion presented before and assume that the MSE between the discretized samples $\tilde{u}_m$ and the estimates $\check{u}_m^{\Delta}(\mathbf{w})$ is to be minimized. The resulting MSE distortion is then given by

$$\check{d}^{\Delta} = E\{||\tilde{\mathbf{U}} - \check{\mathbf{U}}^{\Delta}(\mathbf{W})||^2\}. \tag{11}$$

By applying this relaxation, we are able to formulate a decoding rule solely based on discrete probabilities, such that

$$\check{u}_m^{\Delta}(\mathbf{w}) = E\{\tilde{U}_m|\mathbf{w}\} = \sum_{\forall l \in \mathcal{I}_m} \tilde{u}_m(l) \cdot p(i_m = l|\mathbf{w}). \tag{12}$$

Assuming, that the decoder has knowledge about the mapping functions $f_m(\cdot)$, the required probabilities $p(i_m|\mathbf{w})$ can be easily obtained from the probabilities $p(\mathbf{i})$. Therefore, the probabilities $p(\mathbf{i})$ can be seen as the basis for the conditional mean estimation performed here.

### D. Mismatched Discrete Decoder

Again, the probabilities $p(\mathbf{i})$ might not be available at the decoder, such that the decoder has to perform a conditional mean estimation based on a mismatched probabilities $\hat{p}(\mathbf{i})$, which differ from $p(\mathbf{i})$. The estimates resulting from the mismatched decoding are denoted by $\hat{u}_m^{\Delta}(\mathbf{w})$ and the MSE distortion for mismatched decoding is given by

$$\hat{d}^{\Delta} = E\{||\tilde{\mathbf{U}} - \hat{\mathbf{U}}^{\Delta}(\mathbf{W})||^2\}. \tag{13}$$

Similarly to (12) the estimates $\hat{u}_m^{\Delta}(\mathbf{w})$ can be calculated by

$$\hat{u}_m^{\Delta}(\mathbf{w}) = \sum_{\forall l \in \mathcal{I}_m} \tilde{u}_m(l) \cdot \hat{p}(i_m = l|\mathbf{w}). \tag{14}$$

## IV. Kullback-Leibler Distance

To quantify the mismatch between two PDF's $p(\mathbf{u})$ and $\hat{p}(\mathbf{u})$ we propose the Kullback-Leibler distance (KLD) [1] measured in bits:

$$D(p(\mathbf{u})||\hat{p}(\mathbf{u})) = \int p(\mathbf{u}) \log_2 \frac{p(\mathbf{u})}{\hat{p}(\mathbf{u})} \; d\mathbf{u}. \tag{15}$$

By studying the KLD between the actual and the approximate distribution, we can get some idea of how close our approximation comes to the actual distribution. The smaller the KLD is, the closer we are to the actual probability distribution of the source.

Note, that the expression given in (15) only applies to the continuous probability distributions $p(\mathbf{u})$ and $\hat{p}(\mathbf{u})$, even though the discrete approximation $\hat{p}(\mathbf{i})$ is used for decoding. The continuous probability distributions are used for the optimization because (15) is easier to calculate for this case. However, we can find in [1, p. 231] that the mutual information $I(X;Y)$ between two continuous random variables $X$ and $Y$ is approximately the same as the mutual information $I(X^{\Delta};Y^{\Delta})$ between their quantized versions $X^{\Delta}$ and $Y^{\Delta}$. As the mutual information and the KLD are related measures ($I(X;Y) = D(p(x,y)||p(x) \cdot p(y))$), we conjecture that the KLD $D(p(\mathbf{u})||\hat{p}(\mathbf{u}))$ between the continuous distributions is a close approximation to the KLD $D(p(\mathbf{i})||\hat{p}(\mathbf{i}))$ between the quantized versions.

## V. Problem Statement

Based on the decoder definitions in Section III and the definition of the KLD in Section IV, we formulate three closely related problems:

### A. Problem 1

We are interested in a characterization of the relationship between the MSE distortion $\hat{d}$ of mismatched optimal decoding and the MSE distortion $\check{d}$ of optimal decoding, depending on the KLD $D(p(\mathbf{u})||\hat{p}(\mathbf{u}))$ between the underlying PDF's. It would be most desireable to obtain a closed functional expression $f(\cdot)$ of the form

$$\frac{\hat{d}}{\check{d}} = f\left(D(p(\mathbf{u})||\hat{p}(\mathbf{u}))\right). \tag{16}$$

In case no closed expression for $f(\cdot)$ can be derived, an upper bound for $\hat{d}/\check{d}$ or statements regarding the monotonic behavior of $\hat{d}/\check{d}$ would suffice for many applications.

### B. Problem 2

Based on the observations in Section IV, we would be very interested in a characterization of the relationship between the KLD's $D(p(\mathbf{i})||\hat{p}(\mathbf{i}))$ and $D(p(\mathbf{u})||\hat{p}(\mathbf{u}))$, where the discrete PDF's $p(\mathbf{i})$ and $\hat{p}(\mathbf{i})$ directely result from the continuous PDF's $p(\mathbf{u})$ and $\hat{p}(\mathbf{u})$ by performing PDF optimized quantization as described in Section II, such that $i_m = q_m(u_m)$ for $m = 1, 2, \cdots, M$. It would be desireable to obtain a closed functional expression $g(\cdot)$ of the form

$$D(p(\mathbf{i})||\hat{p}(\mathbf{i})) = g\left(D(p(\mathbf{u})||\hat{p}(\mathbf{u}))\right), \tag{17}$$

or to obtain an upper bound quantifying the discrepancy between both KLD's.

### C. Problem 3

Closely related to Problem 1 and Problem 2, we are interested in a characterization of the relationship between the MSE distortion $\hat{d}^{\Delta}$ of mismatched discrete decoding and the MSE distortion $\check{d}^{\Delta}$ of discrete decoding, depending on the KLD $D(p(\mathbf{i})||\hat{p}(\mathbf{i}))$ between the underlying PDF's. It would be most desireable to obtain a closed functional expression $h(\cdot)$ of the form

$$\frac{\hat{d}^{\Delta}}{\check{d}^{\Delta}} = h\left(D(p(\mathbf{i})||\hat{p}(\mathbf{i}))\right). \tag{18}$$

In case no closed expression for $h(\cdot)$ can be derived, an upper bound for $\hat{d}^{\Delta}/\check{d}^{\Delta}$ or statements regarding the monotonic behavior of $\hat{d}^{\Delta}/\check{d}^{\Delta}$ would be sufficient for many applications.

## References


[1] T. M. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., 1991.

[2] N. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice Hall, 1984.

[3] H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, 1994.

[4] David L. Neuhoff Robert M. Gray. Quantization. *IEEE Trans. on Inform. Theory*, 44, 6 1998.

[5] M. Tüchler, J. Barros, and C. Hausl. Scalable decoding on factor trees: A practical solution for wireless sensor networks. *IEEE Trans. Communications*, 6 2005.