

Relative Entropy and Quantizer Mismatch

R.M. Gray

Department of Electrical Engineering
Stanford University
Stanford, CA 94305

T. Linder

Department of Mathematics and Statistics
Queens University
Kingston, Ontario, Canada K7L 3N6

Abstract

Mismatch in vector quantization or source coding is a measure of the performance loss resulting when a code optimized for one source is applied to another. In lossless compression mismatch is measured by the relative entropy between the source for which the code is designed and the source to which it is actually applied. Recent results in high rate vector quantization theory extend these ideas to the relative entropy between continuous distributions as a measure of quantizer mismatch. We describe the results and some of their implications and examples, including the use of relative entropy in a distortion measure for Lloyd clustering for the design of Gauss mixture models and classified vector quantizers.

1 Introduction

Mismatch in information theory and coding results when a code is optimized for a specific source or distribution, but then applied to another distribution. Perhaps the oldest mismatch result is that for lossless coding. A uniquely decodable lossless code must have a collection of codeword lengths $\ell(i)$ in nats that satisfies the Kraft inequality,

$$\sum_i e^{-\ell(i)} \leq 1. \quad (1)$$

If a discrete source has pmf $p = \{p_i\}$ with Shannon entropy

$$H(p) = - \sum_i p_i \ln p_i,$$

then the smallest average length or *rate* in nats for any lossless uniquely decodable code is easily shown by the divergence inequality to be bound below as $\sum_i p_i \ell(i) \geq H(p)$ with equality if the codeword lengths are chosen as $\ell(i) = -\ln p_i$. As is common, we ignore the constraint of integer word lengths since the unconstrained result also provides a bound and often provides a close approximation to the constrained result. If now this optimal code for the pmf for p is applied to another pmf q , then the resulting average

length is

$$\begin{aligned} \sum_i \ell(i) q_i &= - \sum_i q_i \ln p_i = H(q) + \sum_i q_i \ln \frac{q_i}{p_i} \\ &= H(q) + I(q||p), \end{aligned}$$

where $I(q||p)$ is the *relative entropy* or Kullback-Leibler discrimination of the pmf q with respect to the pmf p . Thus $I(q||p)$ quantifies the performance loss from the optimal performance for q when a code optimized for p is applied to q .

We here describe a recent extension of this result to lossy coding, specifically to high rate entropy constrained vector quantization. Of particular potential interest to image processing applications, we show how this lossy coding mismatch result suggests an approach to clustering Gauss mixture models and the use of such models in compression and classification/segmentation.

2 Preliminaries

The required preliminaries follow [1]. (Ω, \mathcal{B}) is the measurable space consisting of the k -dimensional Euclidean space $\Omega = \mathbb{R}^k$ and its Borel subsets. Assume that X is random vector with a distribution P_f , which is absolutely continuous with respect to the Lebesgue measure V and hence possesses a probability density function (pdf) $f = dP_f/dV$. We assume that the differential entropy $h(f) \triangleq - \int dx f(x) \ln f(x)$ exists and is finite.

A vector quantizer Q can be described by the following mappings and sets:

- An *encoder* $\alpha : \Omega \rightarrow \mathcal{I}$, where \mathcal{I} is a countable index set, say $\{0, 1, 2, \dots\}$, and an associated measurable partition $\mathcal{S} = \{S_i; i \in \mathcal{I}\}$ such that $\alpha(x) = i$ if $x \in S_i$.
- A *decoder* $\beta : \mathcal{I} \rightarrow \Omega$ and an associated reproduction codebook $\mathcal{C} = \{\beta(i); i \in \mathcal{I}\}$. Without loss of generality we assume that the codevectors $\beta(i); i \in \mathcal{I}$ are all distinct.
- A *length function* ℓ of a uniquely decodable lossless index coder satisfying (1). A set of codelengths $\ell(i)$ is said to be *admissible* if (1) holds.

We abbreviate the overall quantizer as q : $q(x) = \beta(\alpha(x))$.

The instantaneous rate of a quantizer is defined by $r(\alpha(x)) = \ell(\alpha(x))$. The average rate is $R_f(Q) = R_f(\alpha, \ell) = E_f r(\alpha(X)) = \sum_i p_i \ell(i)$, where $p_i = P_f(S_i)$ is assumed strictly positive.

Given a quantizer Q , the entropy of the quantizer is defined in the usual fashion by

$$H_f(Q) = H_f(\alpha) = - \sum_i p_i \ln p_i,$$

For any admissible length function ℓ the divergence inequality implies that $R_f(Q) \geq H_f(Q)$ with equality if and only if $\ell(i) = -\ln p_i$. Thus in particular

$$H_f(Q) = \inf_{\ell \in \mathcal{A}} R_f(\alpha, \ell). \quad (2)$$

We assume a distortion measure $d(x, \hat{x}) \geq 0$ and measure performance by average distortion $D_f(Q) = D_f(\alpha, \beta) = E d(X, \beta(\alpha(X)))$. In particular we initially assume squared error distortion $d(x, \hat{x}) = \|x - \hat{x}\|^2 = \sum_{i=1}^k |x_i - \hat{x}_i|^2$ for $x = (x_1, \dots, x_k)$.

The Lagrangian formulation of variable rate vector quantization [3] defines for each value of a Lagrangian multiplier $\lambda > 0$ a Lagrangian distortion $\rho_\lambda(x, i) = d(x, \beta(i)) + \lambda \ell(i)$ and corresponding performance $\rho(f, \lambda, Q) = D_f(Q) + \lambda R_f(Q)$ and an optimal performance $\rho(f, \lambda) = \inf_Q \rho(f, \lambda, Q)$ where the infimum is over all quantizers $Q = (\alpha, \beta, \ell)$ with ℓ admissible. The Lagrangian formulation yields Lloyd optimality conditions for vector quantizers, that is, a necessary condition for optimality is that each of the three components of the quantizer be optimal for the other two:

- For a given decoder β and length function ℓ , the optimal encoder is $\alpha(x) = \operatorname{argmin}_i (d(x, \beta(i)) + \lambda \ell(i))$ (ties are broken arbitrarily).
- The optimal decoder for a given encoder and length function is the usual Lloyd centroid $\beta(i) = \operatorname{argmin}_y E(d(X, y) | \alpha(X) = i)$.
- The optimal length function for the given encoder and decoder is $\ell(i) = -\ln p_i$.

Note that smaller values of λ correspond to higher rates. The next result characterizes the asymptotic performance of optimal entropy constrained vector quantizers as $\lambda \rightarrow 0$, i.e., for asymptotically high rates.

Theorem 1 [1]. *Assume that the distribution P_f of X is absolutely continuous with respect to Lebesgue measure V with pdf $f = dP_f/dV$, that the differential entropy $h(f)$ exists and is finite, and that $H_f(Q_1) < \infty$, where Q_1 is a uniform scalar quantizer with binsize*

1. Then

$$\lim_{\lambda \rightarrow 0} \left(\frac{\rho(f, \lambda)}{\lambda} + \frac{k}{2} \ln \lambda \right) = h(f) + \theta_k \quad (3)$$

where the finite constant θ_k is defined by

$$\theta_k \triangleq \inf_{\lambda > 0} \left(\frac{\rho(u_1, \lambda)}{\lambda} + \frac{k}{2} \ln \lambda \right) \quad (4)$$

and u_1 is the uniform pdf on the k -dimensional unit cube $[0, 1)^k$.

In particular, the limiting constant θ_k depends only on the dimension and not on the pdf.

Define

$$\theta(f, \lambda, Q) \triangleq \frac{E_f d(X, q(X))}{\lambda} + E_f \ell(\alpha(X)) + \frac{k}{2} \ln \lambda - h(f) \quad (5)$$

so that the theorem states that under suitable conditions

$$\lim_{\lambda \rightarrow 0} \inf_Q \theta(f, \lambda, Q) = \theta_k. \quad (6)$$

If one or more of the components is optimized, then it is dropped from the argument of θ , e.g.,

$$\theta(f, \lambda, \alpha, \beta) = \inf_{\ell} \theta(f, \lambda, \alpha, \beta, \ell) \quad (7)$$

$$\theta(f, \lambda) = \inf_{\alpha, \beta, \ell} \theta(f, \lambda, \alpha, \beta, \ell). \quad (8)$$

With this notation the theorem statement becomes

$$\lim_{\lambda \rightarrow 0} \theta(f, \lambda) = \theta_k. \quad (9)$$

The theorem guarantees that if a pdf f satisfies the conditions of the theorem, then there is an *asymptotically optimal* sequence of quantizers q_n for f in the sense that for any decreasing sequence λ_n converging to 0 there exists a sequence of quantizers q_n such that

$$\lim_{n \rightarrow \infty} \theta(f, \lambda_n, q_n) = \theta_k. \quad (10)$$

Given two probability measures P and G on (Ω, \mathcal{B}) for which $P \ll G$ (i.e., P is absolutely continuous with respect to G) and both have densities f and g , then the relative entropy or Kullback-Leibler number is given by

$$I(P||G) = I(f||g) = \int dx f(x) \ln \frac{f(x)}{g(x)}$$

where we have abbreviated the notation to emphasize the dependence on the densities. See, e.g., [4, 5].

3 Quantizer Mismatch

The following high rate variable-rate quantizer mismatch theorem is proved in [2]

Theorem 2 (The mismatch theorem) *Suppose that a probability measure P_g on \mathfrak{R}^k satisfies the conditions of Theorem 1 and has pdf g . Suppose that $Q_n = (q_n, \ell_n)$ is an asymptotically optimal sequence of variable-rate quantizers for P_g , where ℓ_n is the optimal length function for P_g and q_n . Suppose also that $P_f \ll P_g$ and that $dP_f/dP_g = f/g$ is bounded. Then*

$$\lim_{n \rightarrow \infty} \theta(f, \lambda_n, Q_n) = \theta_k + I(f||g). \quad (11)$$

The theorem provides a characterization of the *mismatch* resulting from applying an asymptotically optimal quantizer sequence for one pdf to another: the mismatch is exactly the relative entropy of the mismatched pdf to the design pdf, a continuous analog to the mismatch formula arising in noiseless coding. The result provides a new interpretation of relative entropy as a measure of mismatch for high rate fixed dimension lossy data compression.

4 Gaussian Distributions and Gauss Mixtures

Consider a nonsingular Gaussian pdf $g(x) =$

$$\mathcal{N}(x, \mu, K) \triangleq \frac{1}{(2\pi)^{\frac{k}{2}} |K|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^t K^{-1}(x - \mu)\right)$$

where $\mu = EX$, $K = E[(X - EX)(X - EX)^t]$ is the $k \times k$ covariance matrix, and $|K|$ the determinant of k . The differential entropy is

$$h(g) = - \int dx f(x) \ln f(x) = \frac{1}{2} \ln(2\pi e)^k |K| \quad (12)$$

and it is well known that this differential entropy is the maximum possible over all pdf's corresponding to random vectors with covariance K . This in turn implies that if a sequence $\{\lambda_n, Q_n\}$ is asymptotically optimal for g , then for any pdf f with covariance K for which f/g is bounded, this sequence will yield performance

$$\lim_{n \rightarrow \infty} \theta(f, \lambda_n, Q_n) = \theta_k + \frac{1}{2} \ln(2\pi e)^k |K| - h(f).$$

If all that is known about the pdf is its covariance, designing a code for a Gaussian pdf with the same K will result in a robust code in the sense that the performance is known, but suboptimal in that it is $I(f||g)$ worse than it would have been if the true pdf had been used to design the code. This provides a high rate analog to the Shannon rate distortion results of Sakrison [6] and Lapidoth [7] that an approximately

optimal code designed for a large dimensional independent and identically distributed (i.i.d.) Gaussian vector will yield roughly the same performance on any other i.i.d. vector. Here high rate replaces the assumptions of memorylessness and large dimension.

5 Composite Codes and Lloyd Clustering of Densities

A problem with choosing a worst case pdf to provide a robust quantizer sequence subject to some assumed constraint (e.g., covariance) is that it can be too conservative. One might instead use a collection of models. Each model in the collection could yield a code that was robust for some conditional behavior of the source.

As before let f be the “true” pdf and suppose that Ω_f is its support (which might be all of \mathfrak{R}^k). Assume that $\mathcal{S} = \{S_m; m \in \mathcal{J}\}$, where $\mathcal{J} = \{1, \dots, M\}$, is a finite partition of Ω_f and that $P_f(S_m) > 0$ for all m . Assume also that we have a collection of model pdf's $\{g_m; m \in \mathcal{J}\}$ on \mathfrak{R}^k . We assume further that we have an asymptotically optimal sequence of quantizers for each of the “design” pdf's g_m , that is, for a common decreasing sequence $\lambda_n \rightarrow 0$ we have for each m a quantizer sequence $Q_{m,n}$; $n = 1, 2, \dots$ for which $\lim_{n \rightarrow \infty} \theta(g_m, \lambda_n, Q_{m,n}) = \theta_k$.

Let $Q_n = (\alpha_n, \beta_n, \ell_n)$ be the composite quantizer constructed from the $Q_{m,n} = (\alpha_{m,n}, \beta_{m,n}, \ell_{m,n})$, the partition \mathcal{S} , and a component length function L , that is,

$$\begin{aligned} \alpha_n(x) &= (m, \alpha_{m,n}(x)) \text{ if } x \in S_m \\ \beta_n(m, i) &= \beta_{m,n}(i) \\ \ell_n(m, i) &= L(m) + \ell_{m,n}(i). \end{aligned}$$

Consider the performance resulting when the composite quantizer Q_n is applied to the pdf f . Letting $w_m = P_f(S_m)$, $f_m(x) = f(x)/w_m$ if $x \in S_m$ and 0 otherwise, and choosing index coders optimally yields with some algebra

$$\theta(f, \lambda_n, Q_n) = \sum_m w_m \theta(f_m, \lambda_n, Q_{m,n}).$$

If f/g_m is bounded for each $m = 1, \dots, M$, then we can apply the mismatch theorem to each component to obtain the asymptotic high rate performance

$$\lim_{n \rightarrow \infty} \theta(f, \lambda_n, Q_n) = \theta_k + \sum_m w_m I(f_m||g_m). \quad (13)$$

This equation can be viewed as an extension of the mismatch theorem to composite quantizers. Now the strategy is to divide and conquer: suppose that instead

of a single Gaussian worst case, we are allowed to find a collection $\mathcal{G} = \{g_m; m \in \mathcal{J}\}$ of pdf's from an allowed collection \mathcal{M} of Gaussian pdf's and a partition $\mathcal{S} = \{S_m; m \in \mathcal{J}\}$ of \mathbb{R}^k for use in a composite quantizer. What is the best way to do so? Specifically, for a fixed pdf f and model class \mathcal{M} , find a countable partition \mathcal{S} and model codebook \mathcal{G} which minimizes the mismatch:

$$\bar{I}_f \triangleq \inf_{\mathcal{S}, \mathcal{G}} \bar{I}_f(\mathcal{S}, \mathcal{G}) \quad (14)$$

where

$$\bar{I}_f(\mathcal{S}, \mathcal{G}) = \sum_m P_f(S_m) I(f_m || g_m). \quad (15)$$

This is immediately recognizable as a clustering or quantization problem with reproduction alphabet \mathcal{M} , encoder $a : \mathbb{R}^k \rightarrow \mathcal{J}$ described by the partition $\mathcal{S} = \{S_m\}$ by $a(x) = m$ if $x \in S_m$, $m \in \mathcal{J}$, and decoder $b : \mathcal{J} \rightarrow \mathcal{M}$ defined by $b(m) = g_m$. Thus minimizing mismatch can be viewed as quantizer optimization, finding the encoder/decoder combination minimizing $\bar{I}_f(\mathcal{S}, \mathcal{G})$.

The Lloyd decoder optimization is obvious in this context, given an encoder index m corresponding to encoder cell S_m , the best possible g_m is

$$g_m = \operatorname{argmin}_{g \in \mathcal{M}} I(f_m || g), \quad (16)$$

if the minimum exists. Optimizing the decoder yields another statement of the minimum mismatch problem:

$$\bar{I}_f = \inf_{\mathcal{S}} \sum_m P_f(S_m) \min_{g \in \mathcal{M}} I(f_m || g). \quad (17)$$

To describe a quantizer encoder requires a distortion measure which describes the distortion, say $d_I(x, m)$, between an input vector $x \in \mathbb{R}^k$ and an encoder output (which in turn will produce the reproduction $g_m = b(m)$). The average distortion with respect to the encoder should yield the mismatch, which we are attempting to minimize. As will be shown, this is almost accomplished by defining the "distortion"

$$d_I(x, m) = \ln \frac{f(x)}{g_m(x)} + L(m) \quad (18)$$

where L is an admissible length function which can be optimized along with the encoder and decoder. The first term involves only the shape of the model pdf and it has been used in clustering with the name of a "maximum likelihood" or ML distortion since minimizing this distortion over m for a given x is equivalent to choosing the maximum likelihood estimate for

m assuming the vector was produced by one of the models g_m [9]. It is equivalent to the quadratic discrimination analysis (QDA) decision rule in statistics. Using a Lagrangian formulation of variable rate vector quantization as in [9] results in the addition of a $\nu L(m)$ term to the ML distortion for a Lagrange multiplier ν , so the d_I distortion can be viewed as such a distortion for the special case of $\nu=1$. The problem of course is that this is not a true distortion since it need not be nonnegative, but the average *is* nonnegative.

The $L(m)$ term represents a cost or penalty for choosing the index m and hence can be considered as a constraint on the encoder partition. We require as before that it be an admissible length function and satisfy 1. Once a distortion measure is specified, the optimal encoder must be a minimum distortion encoder and hence for a given decoder codebook \mathcal{G}

$$a(x) = \operatorname{argmin}_m d_I(x, m) = \operatorname{argmin}_m (L(m) - \ln g_m(x)).$$

In particular, the encoder does not involve knowledge of f except indirectly through the w_m . The corresponding encoder partition \mathcal{S} will then yield average distortion

$$\int dx f(x) d_I(x, a(x)) \geq \sum_m w_m I(f_m || g_m)$$

with equality if and only if we choose the optimal length function $L(m) = -\ln w_m$. Thus if we choose an optimal decoder and length function for a partition, the average distortion according to d_I is exactly the mismatch. Iterating the Lloyd optimality properties of optimizing encoder, decoder, and length function can only decrease average distortion and hence also the mismatch.

The Lloyd algorithm for minimizing mismatch produces a collection of models $g_m \in \mathcal{M}$ drawn from some set \mathcal{M} together with a probability mass function, w_m . A collection of pdf's together with a pmf can be viewed as a *mixture* and hence the proposed algorithm can be viewed as a means of fitting mixtures of specified families of densities to an arbitrary pdf.

Gauss Mixture Codes

Let \mathcal{M} consist of all nonsingular Gaussian pdf's. Again begin by considering the centroid $g \in \mathcal{M}$ as the Gaussian pdf g minimizing $I(f || g)$. This is accomplished by some algebraic manipulation using relative entropies for Gaussian pdf's as found, e.g., in Kullback [4]. In particular, given a Gaussian pdf g with mean μ and covariance K and a pdf f with mean μ_f and covariance K_f . Then

$$I(f || g) = -h(f) + \frac{1}{2} \ln(2\pi e)^k |K_f| + \left[\frac{1}{2} \ln \frac{|K|}{|K_f|} + \right.$$

$$\frac{1}{2} \text{Trace}(K_f K^{-1}) - \frac{k}{2} + \frac{1}{2} (\mu - \mu_f)^t K^{-1} (\mu - \mu_f).$$

The bracketed term is exactly the relative entropy between a Gaussian pdf with mean μ_f and covariance K_f and a second Gaussian pdf with mean μ and covariance K (e.g., p. 189 of Kullback [4]). Thus, in particular, the quantity is nonnegative and will in fact be zero with the choices $\mu_f = \mu$ and $K_f = K$, i.e., if we choose the mean and covariance of the model g to match the mean and covariance of f . The rightmost term is nonnegative and will also be 0 if $\mu_f = \mu$. With these choices we are left with

$$I(f||g) = -h(f) + \frac{1}{2} \ln(2\pi e)^k |K_f|$$

and the centroid g is the Gaussian which has as mean and covariance the mean and covariance with respect to f .

Again consider the conditional relative entropy arising with a composite quantizer. Given an encoder partition \mathcal{S} , the centroids are given as above with f replaced by f_m : define the conditional mean $\mu_{f_m} = E_{f_m} X$ and the conditional covariance $K_{f_m} = E_{f_m} [(X - \mu_{f_m})(X - \mu_{f_m})^t]$ (conditioned on $X \in S_m$). Then

$$b(m) = g_m = \underset{g \in \mathcal{M}}{\text{argmin}} I(f_m || g) = \mathcal{N}(x, \mu_{f_m}, K_{f_m}).$$

The encoder depends on the pdf f_m *only through the mean and covariance*.

For a model quantizer with an optimal decoder, the mismatch can be expressed simply as

$$\sum_m w_m I(f_m || g_m) = -h(f) + H(Z) + \frac{1}{2} \sum_m w_m \ln(2\pi e)^k |K_{f_m}|$$

where Z is a discrete random variable with pmf $P(Z = m) = w_m = P_f(S_m)$ for all m . The average distortion forces a balance between the rightmost term, which tries to match Gaussian models to partition cells, and the entropy term, which puts a cost on partition cells.

When using individual Gaussian models with optimal codebooks and length functions, the distortion $d_I(x, m)$ becomes $\ln f(x) - \ln w_m + \frac{1}{2} \ln((2\pi)^k |K_{f_m}|) + \frac{1}{2} (x - \mu_{f_m})^t K_{f_m}^{-1} (x - \mu_{f_m})$ and the optimal minimum distortion encoder picks m to minimize this distortion given an observed x .

The Gauss mixture design can be used to design classifiers and hence image segmentation algorithms in several ways. A Gauss mixture can be designed for

a complete source and used to design a classified VQ. Each index can be labeled by the maximum a posteriori class for that index as estimated from the training set. Alternatively, a separate Gauss mixture can be designed for each class and a separate VQ codebook designed for each mixture. A new vector (or an entire image) is encoded using the collection of codebooks and the codebook with the smallest average distortion indicates the class. This is an extension to images of a codebook approach to speech recognition [8]. Experimental examples will be presented in the talk and further examples can be found in [10, 9].

References

- [1] R. M. Gray, T. Linder, and J. Li, "A Lagrangian formulation of Zador's entropy-constrained quantization theorem," *IEEE Trans. Inform. Theory*, pp. 695–707, vol. 48, Mar. 2002.
- [2] R. M. Gray and T. Linder, "Mismatch in high rate entropy-constrained vector quantization," submitted for publication. Preprint available at <http://ee.stanford.edu/~gray/mismatch.pdf>.
- [3] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust. Speech and Signal Proc.*, vol. 37, pp. 31–42, Jan. 1989.
- [4] S. Kullback. *Information Theory and Statistics*, Dover, New York, 1968.
- [5] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, 1990.
- [6] D. J. Sakrison, "Worst sources and robust codes for difference distortion measures," *IEEE Trans. Inform. Theory*, vol. 21, pp. 301–309, May 1975.
- [7] A. Lapidoth, "On the role of mismatch in rate distortion theory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 38–47, Jan. 1997.
- [8] J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 473–491, Jul. 1983.
- [9] R.M. Gray, J.C. Young, and A. K. Aiyer, "Minimum discrimination information clustering: modeling and quantization with Gauss mixtures," *Proceedings 2001 IEEE International Conference on Image Processing*, pp. 14–17, Thessaloniki, Greece, October 2001.
- [10] R. M. Gray, "Gauss mixture vector quantization," *Proceedings 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1769–1772, Salt Lake City, May 2001.