Rate Distortion Behavior of Sparse Sources

Claudio Weidmann, Martin Vetterli

Rate Distortion Behavior of Sparse Sources

Claudio Weidmann, *Member, IEEE*, and Martin Vetterli, *Fellow, IEEE*

***Abstract*—The rate distortion behavior of sparse memoryless sources is studied. These serve as models of sparse signal representations and facilitate the performance analysis of "sparsifying" transforms like the wavelet transform, and nonlinear approximation schemes. For strictly sparse binary sources with Hamming distortion, $R(D)$ is shown to be almost linear. For non-strictly sparse continuous-valued sources, termed compressible, two measures of compressibility are introduced: incomplete moments and geometric mean. The former lead to low- and high-rate upper bounds on mean squared error $D(R)$, while the latter yields lower and upper bounds on source entropy, thereby characterizing asymptotic $R(D)$ behavior. Thus the notion of compressibility is quantitatively connected with actual lossy compression. These bounding techniques are applied to two source models: Gaussian mixtures and power laws matching the approximately scale-invariant decay of wavelet coefficients. The former are versatile models for sparse data, which in particular allow to bound high-rate compression performance of a scalar mixture compared to a corresponding unmixed transform coding system. Such a comparison is interesting for transforms with known coefficient decay, but unknown coefficient ordering, e.g. when positions of highest-variance coefficients are unknown. The use of these models and results in distributed coding and compressed sensing scenarios is also discussed.**

***Index Terms*—Sparse signal representations, rate distortion theory, memoryless systems, entropy, transform coding.**

I. Introduction

SPARSE signal representations are the basis of state-of-the-art lossy compression and applied compressive sampling / compressed sensing. The fundamental appeal of sparsity lies in the property that a small number of coefficients carries the bulk of the signal energy, or more generally the part of the signal that is relevant to the application, e.g. perceptually. In the case of traditional lossy compression, sparsity provides a first stage of compression by reducing the number of coefficients needed for approximate reconstruction (by nonlinear approximation) [1]. In the case of sparse sampling [2] and compressed sensing [3], [4], sparsity enables sampling a signal below its apparent Nyquist rate, while incurring a minimal increase in distortion. This is achieved by "universally" sampling the signal (e.g. using an appropriate random basis) and imposing a sparsity constraint on its reconstruction. Part of the appeal of such methods comes from the fact that the computational complexity of "sparsifying" transform and nonlinear approximation is moved from the encoder (sampling device) to the decoder (reconstruction device), that is, the encoder is kept "simple" and non-adaptive.

The analysis of both nonlinear approximation (NLA) and compressed sensing (CS) has long focused on the *number of coefficients / samples* required to achieve reconstruction at a given distortion level. However, this ignores the fact that most applications involve some form of digital transmission or storage, which requires quantizing analog continuous-valued coefficients. The approach taken in this paper is to study the *number of bits* needed to achieve a given distortion, by modeling the output of a sparsifying transform as a *sparse source*, whose rate distortion behavior can be characterized. Such analysis has the advantage that it characterizes the ultimate compression trade-off between rate (in bits/sample) and distortion, independently of the scheme under consideration. Under the assumption that the sparsifying transform is known to both encoder and decoder, it does not matter whether the transform is used at the encoder (as in quantized NLA, i.e. adaptive lossy compression) or at the decoder (CS with quantized samples), provided that the ultimate goal is to reconstruct the sparse source signal with the smallest distortion possible for a given bit budget. This means that such information-theoretic analysis does not take into account practical complexity issues, like e.g. encoders with limited processing capabilities, which might favor CS over NLA.

A central aspect of our approach is how to model sparse sources and how to measure their (approximate) sparsity. We focus on simple memoryless models that suffice to gain insights on the relation between sparsity and rate distortion behavior. Wavelet coefficients will serve as a practical example of a sparse source throughout the paper, since the material presented here has its roots in our work on understanding wavelet image compression. Besides this, the wavelet transform is perhaps the best known sparsifying transform, and it also plays a key role in recent CS applications such as the "single pixel camera" [5]. Since unitary transforms (or nearly unitary ones, like the popular 9/7 biorthogonal wavelet) leave vector norms unchanged, for mean squared error (MSE) distortion measure it is sufficient to study the rate distortion function of sparse sources modeling the transform coefficients. The main focus of this paper is thus the characterization of non-strictly sparse continuous-valued sources. We adopt the often-used term *compressible* to denote such sources. The key questions that will be addressed are how to quantitatively measure sparsity and how to relate such measures with the rate distortion properties of a source.

Manuscript received December 20, 2008; revised October 11, 2011. Accepted for publication February 3, 2012. The material in this paper was presented in part at the Data Compression Conference, Snowbird UT, March 1999 and 2000, and at the IEEE International Symposium on Information Theory, Washington DC, June 2001. Part of this work stems from the Ph.D. thesis of the first author and was supported by an ETHZ/EPFL fellowship.

Claudio Weidmann was with the Audiovisual Communications Laboratory, EPFL, Lausanne, Switzerland, and is now with ETIS – ENSEA / Univ Cergy-Pontoise / CNRS UMR 8051, Cergy-Pontoise, France (e-mail: claudio.weidmann@ieee.org).

Martin Vetterli is with the Audiovisual Communications Laboratory, EPFL, Lausanne, Switzerland and with the Department of EECS, UC Berkeley, Berkeley CA 94720.

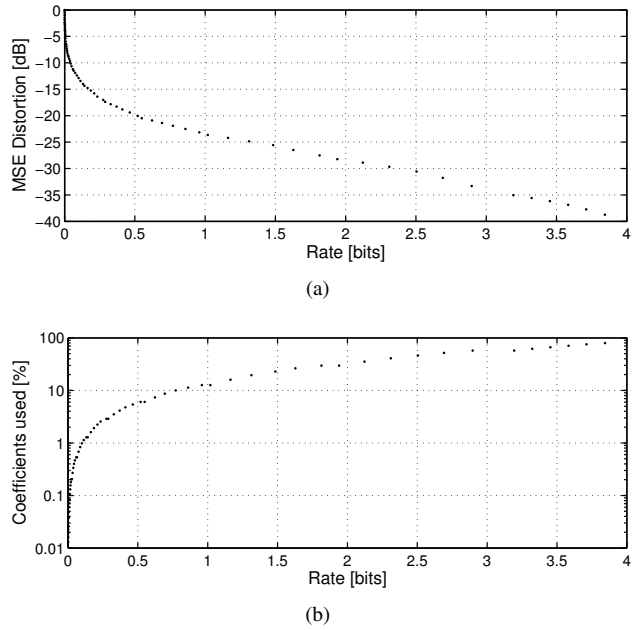


Fig. 1. (a) Operational distortion rate points of a wavelet coder applied to the Lena image. The knee shape, leading from steep decay at low rates to the asymptotic -6 dB/bit slope, is typical for such image coders. (b) At low rates, only a small fraction of coefficients is quantized to nonzero values, all the others are not used in the reconstruction of the image.

Sparsifying transforms, which are the main tool for obtaining sparse signal representations, have been studied in various compression-related settings. For example, the success of wavelet-based coding is often attributed to the ability of wavelets to "isolate" singularities, something Fourier bases fail to do efficiently [1]. Thus, a piecewise smooth signal is mapped through the wavelet transform into a sparse set of non-zero transform coefficients, namely coefficients around discontinuities, as well as coefficients representing the general trend of the signal [6]. While this behavior is well understood in terms of nonlinear approximation power (that is, approximation by the N largest terms of the wavelet transform, see [7] for a thorough treatment), the rate distortion behavior is more open. Early work on NLA of random functions [8] concentrated on approximation error as a function of the number of approximation terms, neglecting the trade-off between the rate needed to identify these terms and the rate used to quantize each term. Mallat and Falzon [9] were the first to analyze the operational low-rate behavior of image transform coding, which is very different from the behavior expected from classic Karhunen-Loève transform (KLT) theory. In essence, at low rates only few wavelet coefficients are involved in the approximation of piecewise smooth functions, leading to a decay of the distortion rate function that is steeper than the classic exponential decay in the case of Gauss-Markov processes and the KLT. This result had been observed experimentally in low-rate image coding; see Fig. 1 for an example.

A key difference between compressing jointly Gaussian processes using the KLT and compressing piecewise smooth processes with the wavelet transform lies in the identification of the set of *significant* coefficients that are quantized and used for reconstruction. In the KLT case, the optimal rate allocation strategy is reverse water-filling [7, Sec. 11.3], [10, Sec. 13.3.3], meaning that statistical signal properties (the eigenvalues of the covariance matrix) determine *a priori* the set of coefficients for a given reconstruction quality. The KLT approximation is "linear" (up to quantization) and non-adaptive, in the sense that two sample vectors with the same covariance matrix will be approximated using the same set of coefficients, which spans a subspace. In the wavelet approach, the approximation is "nonlinear" and adaptive, since the set of coefficients is chosen *a posteriori* based on the transformed signal realization and may thus change from instance to instance. This underlines the importance of coding the positions of the significant coefficients in a sparse vector; see [11] for a thorough analysis in the context of wavelets.

The above-cited results indicate the interest to study the rate distortion behavior of sparse signal representations in more depth, in particular, to narrow down rates and distortions within constants, avoiding the loose factors in the exponent that are often present in approximation results. Recently, sparse sources have also received renewed attention with the work on sparse sampling [2] and compressed sensing (CS) [3], [4]. Their rate distortion behavior is still being studied, with some initial results in [12]. Thus, the present paper fills a gap, giving either precise results or tight bounds on the rate distortion behavior of models that serve as benchmarks for these methods.

The remainder of this paper is organized as follows. Section II presents three classes of source models, namely strictly sparse binary sources and mixed discrete/continuous "spike" sources, as well as non-strictly sparse continuous sources. It also briefly introduces some essential information-theoretic definitions and tools.

Section III looks at strictly sparse sources. Binary vectors with Hamming distortion are studied in Section III-A as a model for coding the positions of a set of coefficients. Closed-form expressions for $R(D)$ are derived for the case when the number of non-zero entries is known; these hold also for non-sparse sources. For sparse binary sources, $R(D)$ is found to be essentially linear. Section III-B then considers a mixed discrete/continuous spike source, in which a Bernoulli (position) source switches a Gaussian (value) source on or off. The MSE distortion rate behavior is characterized using upper bounds. Sparse spikes help explaining the steep distortion decay in low-rate NLA, but they fail to model the behavior at medium to high rates, for which continuous sources are more appropriate.

Section IV opens the main theme by introducing two ways of measuring *compressibility* (non-strict sparsity) of continuous-valued sources: using *incomplete moments* and using the *geometric mean*. Based on incomplete moments, Section V introduces upper bounds on MSE $D(R)$ and applies them to a popular power-law model for approximately scale-invariant data, such as wavelet coefficients. Section VI then presents lower and upper bounds on the source entropy using the geometric mean and the variance, thereby characterizing the asymptotic rate distortion behavior of a source as a function of its compressibility. In fact, these bounds on $D(R)$ and entropy hold for continuous sources with arbitrary

compressibility, i.e. also for non-sparse sources.

The theme of compressible sources is continued in Section VII, which considers Gaussian mixture models, showing that simple two-component mixtures already capture the essential $D(R)$ characteristics (the knee shape) of sparse sources. Based on incomplete moments, a notion akin to classic transform coding gain is introduced. In the case of Gaussian transform coefficients, it is possible to bound the loss in coding gain if the coefficients are randomly mixed (that is, if one knows only their variances, but not their positions).

Finally, Section VIII briefly outlines how the results on compressible sources can be applied to distributed coding and compressed sensing scenarios.

II. Models, Definitions and Tools

A. Models for Sparse Sources

When using sparse signal representations as building blocks for lossy source coding, the goal is to concentrate most of the signal energy in as few coefficients as possible. Lossy compression then proceeds by selecting a subset of coefficients that will be quantized. Nonlinear approximation (NLA) methods will generally select the largest coefficients first, other (linear) methods might select a fixed set depending on the coding rate or some other criterion. The quality of the reconstruction from the quantized coefficients will be measured with an appropriate distortion measure.

The coefficients representing the signal will be modeled as coming from a memoryless *sparse source* X, which emits an i.i.d. sequence of random variables $X_1, X_2, \ldots$. For simplicity, we will use X to denote both the source and the random variable(s) that it emits. Before presenting the different models, we need to clarify the notion of "sparse source." We will say that a source is *strictly sparse* if it emits the value zero with positive probability. Clearly, the closer this probability is to 1, the sparser is the source. A natural measure of sparsity in this situation is the normalized Hamming weight of a sample vector, $\frac{1}{n} w_H(\boldsymbol{x})$, which asymptotically equals $\Pr\{X \neq 0\} < 1$, such that smaller values indicate a sparser source. The Hamming weight is $w_H(\boldsymbol{x}) = d_H(\boldsymbol{x}, \mathbf{0})$, where $d_H(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{n} \sum_{i=1}^{n} d_H(x_i, \hat{x}_i)$ and $d_H(x, \hat{x}) = \mathbb{1}_{x \neq \hat{x}}$ are the Hamming distances between vectors and symbols, respectively.

A more general notion of sparsity will encompass sources that emit sparse sequences affected by weak background noise, which has negligible energy compared to the sparse component. Such *compressible* (non-strictly sparse) sources can be modeled by proper continuous random variables. These have $\Pr\{X \neq 0\} = 1$ and therefore Hamming weight cannot be used to measure their sparsity; alternative measures will be proposed in Section IV.

Three different classes of sparse source models will be studied:

1) *Sparse binary sources* might model a *significance map* or *sparsity pattern*, that is, the binary map recording the positions of *significant* coefficients in an NLA scheme. (These are the coefficients which are actually used to reconstruct the signal.) We will analyze both sources emitting vectors of length N containing exactly K ones and Bernoulli-p (binary memoryless) sources, emitting sequences of i.i.d. binary random variables.
2) *Spike sources* are a generalization of sparse binary sources, where each binary one is associated with a continuous random variable. In particular, we will study the product of a Bernoulli-p source (emitting 0 or 1) and a memoryless Gaussian source, using the MSE distortion measure. This might serve as a crude model of very low rate wavelet-based NLA coding, when only a tiny subset of coefficients is used to represent the signal.
3) *Compressible sources* are memoryless sources emitting i.i.d. continuous random variables with a peaked unimodal density (the mode is assumed to be zero). Examples are power laws, Laplacians and generalized Gaussian densities with exponent smaller than one. Such sources can be used as a first-order model for wavelet coefficients in e.g. image coding [13]. In particular, we will show that very simple *Gaussian mixtures* are sufficient to capture the key aspects of the rate distortion behavior of sparse wavelet coefficients.

B. Definitions and Tools

The rate distortion function $R(D)$ of a source was introduced by Shannon to measure the minimal amount of information rate required to describe the source output within average distortion D [14], [15]. It is the minimal rate needed by an optimal (high-dimensional) vector quantizer, that is by an optimal lossy compressor. This operational definition is found to be equal to the *information* rate distortion function $R^{(I)}(D)$ [10, Theorem 13.2.1],

$$R(D) = R^{(I)}(D) = \min_{f(\hat{x}|x):\ \mathrm{E}\, d(X,\hat{X}) \leq D} I(X;\hat{X}), \quad (1)$$

where $\hat{X}$ is the reconstruction random variable defined via the conditional probability mass (or density) function (pmf or pdf) $f(\hat{x}|x)$, $I(X;\hat{X}) = \mathrm{E} \log \frac{f(X,\hat{X})}{f(X)f(\hat{X})}$ is the mutual information between X and $\hat{X}$, and $d(x,\hat{x})$ is the distortion measure. The expected distortion is obtained over the joint distribution $f(x,\hat{x}) = f(x)f(\hat{x}|x)$. Equality (1) implies that the information-theoretic function $R^{(I)}(D)$ describes the ultimate performance limits of lossy compression (and it also allows us to drop the superscript $^{(I)}$ in the following). These definitions and results hold for discrete and continuous sources, as well as mixed discrete/continuous sources, under appropriate conditions on $f(x)$ and $d(x,\hat{x})$ [16], [10, Chap. 13]. Closed-form expressions for $R(D)$ are known for the binary memoryless (Bernoulli-p) source with Hamming distortion and the Gaussian source with squared error distortion. Alternatively, the distortion rate function $D(R)$, the inverse of $R(D)$, measures the minimal distortion achievable with a given description rate. We use both functions interchangeably, but in figures we always plot distortion over rate.

This work considers only memoryless sources and single-letter distortion measures. For the first of the above models, sparse binary sources, a natural distortion measure is the Hamming distance $d_H(x,\hat{x})$. The other two models, spike

sources and compressible sources, output continuous values and correspond to situations where signal energy will be measured with the square norm. Hence we will use the mean squared error (MSE) as distortion measure, corresponding to $d(x, \hat{x}) = (x - \hat{x})^2$.

Two key results on continuous sources that will be used as tools throughout the paper are as follows. The *Gaussian upper bound* states that the MSE distortion rate function of a memoryless continuous source with variance σ^2 is upper-bounded by the distortion rate function of a Gaussian source with the same variance [14, Theorem 23],

$$D(R) \leq \sigma^2 e^{-2R}. \tag{2}$$

Note that in this work, all rates are expressed in nats and all logarithms are natural, unless otherwise stated. The *Shannon lower bound* (SLB) states that the MSE rate distortion function of a memoryless continuous source is lower-bounded by that of a Gaussian source with the same differential entropy [14, Theorem 23],

$$R(D) \geq R_{\mathrm{SLB}}(D) = h(X) - \frac{1}{2}\log(2\pi e D), \tag{3}$$

where $h(X) = -\operatorname{E}\log f(X)$ is the differential entropy of the source. For a large class of sources with sufficiently "nice" densities, the MSE SLB (3) is asymptotically tight for small distortions (large rates), that is $R(D) - R_{\mathrm{SLB}}(D) \to 0$ as $D \to 0$; see e.g. [16, Sec. 4.3.4] or [17]. Thus we will use bounds on the entropy of compressible sources to characterize their asymptotic rate distortion behavior.

III. Strictly Sparse Sources

A. Sparse Binary Sources

We will first study memoryless binary vector sources that emit exactly K ones in a vector of length N, after which we look at the simple scalar binary memoryless (Bernoulli-p) source. Reconstruction fidelity is measured with Hamming distortion, which is equivalent to a frequency of error criterion where both types of errors have the same cost (coding a one when there is none and vice-versa).

Definition 1 *The* binary (K, N) source *is a memoryless source that emits binary vectors of length N and Hamming weight K, with uniform probability over the $\binom{N}{K}$ possible patterns.*

Since the source alphabet size is finite, the rate distortion problem is not a proper vector-valued problem and can actually be solved with the methods for discrete memoryless sources summarized in Appendix A.

a) Binary Vectors of Weight 1: The simplest case of a binary $(1, N)$ source $\boldsymbol{X}$ is equivalent to a memoryless uniform source U with alphabet $\mathcal{U} = \{1, 2, \ldots, N\}$. Using the standard basis vectors $\boldsymbol{e}_i$ we can write $\boldsymbol{X} = \boldsymbol{e}_U$. It can be shown (see Theorem 14 from [16] in Appendix A) that just one additional reconstruction letter is needed to achieve the Hamming rate distortion bound, and it will map to the all-zero vector $\mathbf{0}$. To see that it can only be the all-zero vector, consider the source alphabet $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N\}$, which consists of all vectors of Hamming weight one. Any other non-zero vector will be at Hamming distance one or more from these vectors and thus can only worsen the distortion achieved by the all-zero vector, which is exactly one. If we define $\hat{\mathcal{U}} = \mathcal{U} \cup \{0\}$ and $\boldsymbol{e}_0 = \mathbf{0}$, then everything fits nicely. Using $\hat{u} = 0$ corresponds to not coding the position. We get the distortion measure

$$\rho(u, \hat{u}) = d_H(\boldsymbol{e}_u, \boldsymbol{e}_{\hat{u}}) = d_H(u, \hat{u}) \cdot (1 + d_H(\hat{u}, 0)).$$

Thus "giving the right answer" has zero distortion, a wrong answer two, and not answering costs one distortion unit.

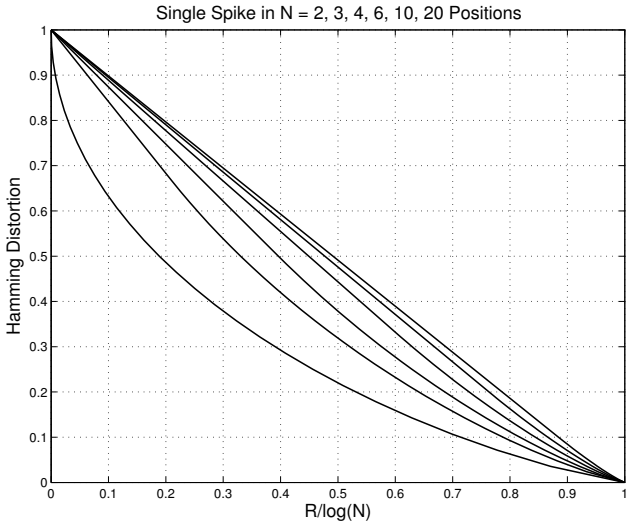


Fig. 2. $D(R)$ for the binary $(1, N)$ source with Hamming distortion, $N = 2 \ldots 20$ (bottom to top curve). The rate has been normalized by $\log N$. For $N \to \infty$, $D(R)$ becomes a straight line, see (4).

Proposition 1 *The Hamming rate distortion function of a binary $(1, N)$ source for $N \geq 2$ is*

$$R(D) = \begin{cases} (1 - D)\log(N - 1), & \frac{2}{N} < D \leq 1, \\ \log N - \frac{D}{2}\log(N - 1) - h_b\left(\frac{D}{2}\right), & 0 \leq D \leq \frac{2}{N}, \end{cases} \tag{4}$$

where $h_b(p) = -p\log p - (1 - p)\log(1 - p)$ is the binary entropy function.

The proof appears in Appendix B; Fig. 2 shows a set of typical $D(R)$ functions. As N becomes large, the linear segment dominates the rate distortion characteristics. In the special case $N = 2$, the solution degrades to twice the $D(R)$ function of a binary symmetric source.

b) Binary Vectors of Weight K: The general binary (K, N) source for $K \geq 2$ emits one of the $\binom{N}{K}$ binary vectors of length N and Hamming weight K, uniformly at random. Its Hamming rate distortion function can be obtained in similar fashion to the $(1, N)$ source, if the additional reconstruction letter is the all-zero vector [18, Sec. 3.2.2]. However, this choice of $\hat{\mathcal{U}}$ is optimal only for low distortions. Determining the best $\hat{\mathcal{U}}$ for higher distortions requires a cumbersome case-by-case analysis that can be avoided in view of the following result for sparse Bernoulli-p sources.

c) Sparse Binary Memoryless Sources: The simplest model of a sparse binary source is a Bernoulli-p binary memoryless source (BMS) with $p = \Pr\{X = 1\} \ll 1$. Its extension to blocks of symbols may be considered as a randomized version of the above binary vector models, since blocks of N samples will contain close to pN ones on average, instead of a fixed number K.

Proposition 2 *Consider a Bernoulli-p source ($p \leq \frac{1}{2}$) with normalized distortion $d = D/p$, where D is Hamming distortion. Then the normalized rate distortion function is asymptotically linear when $p \to 0$:*

$$\lim_{p\to 0} \frac{R(pd)}{h_b(p)} = 1 - d, \quad 0 \leq d \leq 1$$

Proof: The rate distortion function of the BMS is $R(D) = h_b(p) - h_b(D)$ for $D \leq p \leq \frac{1}{2}$ [10, Thm. 13.3.1]. Therefore

$$\begin{aligned}\frac{R(pd)}{h_b(p)} &= 1 - \frac{h_b(pd)}{h_b(p)} \\ &= 1 - \frac{pd\log(pd) + (1-pd)\log(1-pd)}{p\log(p) + (1-p)\log(1-p)},\end{aligned}$$

from which, by applying Bernoulli-de l'Hospital's rule twice,

$$\begin{aligned}\lim_{p\to 0} \frac{R(pd)}{h_b(p)} &= 1 - \lim_{p\to 0} \frac{d\log(pd) - d\log(1-pd)}{\log p - \log(1-p)} \\ &= 1 - \lim_{p\to 0} \frac{d/p + d^2/(1-pd)}{1/p + 1/(1-p)} \\ &= 1 - d\end{aligned}$$

■

Proposition 2 shows that if we normalize the rate and the distortion by their maxima, $h_b(p)$ and p, respectively, the rate distortion function becomes linear for sparse binary sources with $p \to 0$.

d) Remarks: For both the vector model and the BMS the rate distortion function becomes linear for very sparse sources, for which the average Hamming weight approaches zero. The interest of the vector model lies mainly in the fact that it yields analytic expressions for $R(D)$, of which there are not many examples in rate distortion theory.

The consequence of this "almost linear" behavior of sparse binary sources is the following: to encode sequences of length n at intermediate rates $0 < R < nh_b(p)$, it is not necessary to use a complex lossy encoder, but one can simply encode the positions of the ones in sequential fashion using a lossless encoder (e.g. an arithmetic coder), until the bit budget R is used up.

B. Spike Sources

The previous section studied sparse binary sources that may model the position of significant coefficients. Now we also consider the values of those coefficients, by modeling them as continuous random variables. The resulting model is a discrete-time stochastic process that is zero almost all the time, except in a few positions, where *spikes* stick out. Distortion will be measured by the mean squared error (MSE). A simple model of a spike source can be obtained by multiplying the outputs of a binary source (emitting 0 or 1) and a memoryless continuous source. The binary source simply switches the value source on or off. Here we consider only Gaussian-distributed values, because they provide a worst-case benchmark for MSE distortion.

Definition 2 *The* Bernoulli-Gaussian (BG) spike source *emits i.i.d. random variables that are the product of a binary random variable U with $\Pr\{U=1\} = p$ and $\Pr\{U=0\} = 1-p$ and an independent zero-mean Gaussian random variable V with variance σ_v^2. Using Dirac's delta function, the "pdf" of the BG spike can be written as*

$$f(x) = (1-p)\delta(x) + p\frac{1}{\sqrt{2\pi}\sigma_v} e^{-x^2/2\sigma_v^2}. \tag{5}$$

BG spikes are *mixed* random variables that have both a discrete and a continuous component. From (5), it is clear that the distribution function of such random variables is not absolutely continuous in general and therefore most results of standard rate distortion theory do not hold. The spike entropy cannot be computed with the usual integral, but only via mutual information conditioned on the discrete part [19, Ch. 2]. With this method, Rosenthal and Binia [20] derived the asymptotic ($D \to 0$) rate distortion behavior of mixed random variables, as well as certain mixed random vectors. Their result coincides with the simple upper bound (6) presented below if the continuous part is Gaussian, otherwise their result is tighter. Later, György *et al.* [21] extended these asymptotic results to random vectors with more general mixed distributions and to a wide class of sources with memory.

A simple upper bound on $D(R)$ of the spike source can be derived using an adaptive two-step code: 1. all samples with magnitudes above $\epsilon = 0$ are classified as spikes and their positions encoded with a bitmap using $h_b(p)$ nats/sample; 2. the spike values are encoded with a Gaussian random codebook using $\frac{1}{2}\log\frac{p\sigma_v^2}{D}$ nats/spike [10, Thm. 13.3.2]. (A generalization of this coding scheme was shown to be asymptotically optimal in the limit of small distortions for some mixed-distribution sources with memory in [21].) The resulting upper bound expressed as a function of rate is

$$D(R) \leq p\sigma_v^2 \exp\left(-\frac{2R - 2h_b(p)}{p}\right), \qquad R \geq h_b(p). \tag{6}$$

This bound is loose at high distortions (i.e. low rates) and can be improved by coding only a fraction of the spikes. In particular, a tighter bound is obtained by varying the classification threshold ϵ and optimizing over the resulting family of upper bounds; the result will be stated in Section V.

Fig. 3 shows the bound (6) and the optimized low-rate bound (16) from Section V, together with $D(R)$ estimated numerically with the Blahut-Arimoto algorithm [10, Sec. 13.8] for different values of p. The asymptotic distortion decay is on the order of $-\frac{6}{p}$ dB/bit, which can be much steeper than the -6 dB/bit typical of absolutely continuous random variables. This decay behavior is representative of spike sources, regardless whether the value is Gaussian or not.

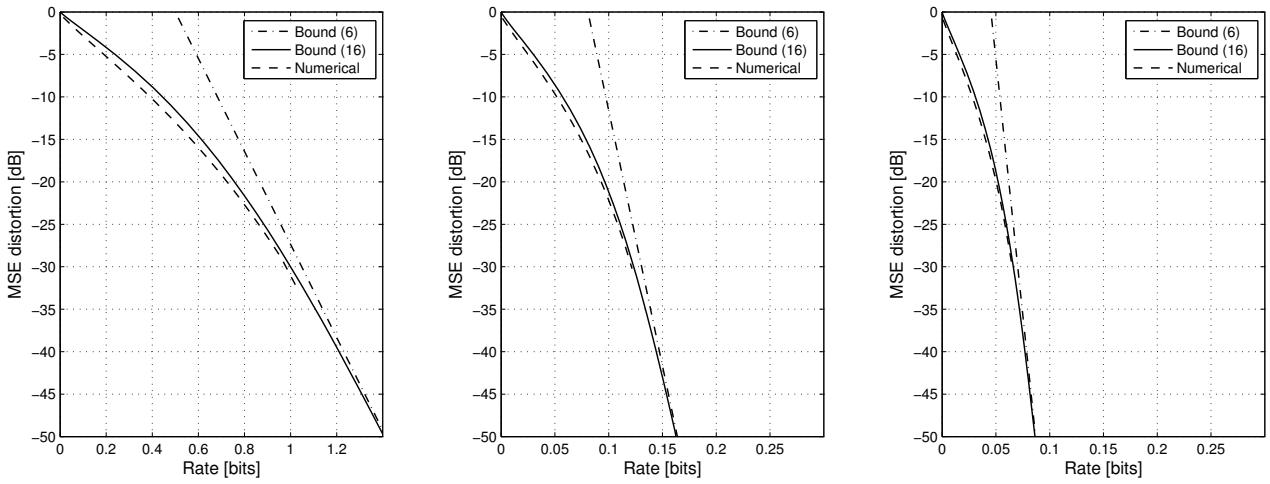


Fig. 3. Distortion rate behavior of Bernoulli-Gaussian spikes for different values of the Bernoulli-p parameter (normalized to unit variance): (a) $p = 0.11$, (b) $p = 0.01$, (c) $p = 0.005$.

Comparing with Fig. 1, we see that the spike $D(R)$ behavior is very different from the one observed in actual lossy compression. Thus the spike source is certainly not a good general model for sparse transform coefficients. However, it explains the steep $D(R)$ decay that can be achieved at very low rates, when only very few coefficients are used to represent the data. When the rate is higher, the spike model fails, because the abrupt change from zero to a Gaussian value distribution does not reflect the coefficient decay actually observed (i.e. the non-strict sparsity that will be considered in the next section). There are other applications of spikes, such as using them as a benchmark for transform coding. In the case of data-independent (linear) rate allocation, any transform of the spike process yields worse performance compared to nonlinear approaches [1]. For constant-value spikes ("1 in N" as in Sec. III-A), any KLT basis contains the vector $[1, 1, \ldots, 1]^T$ and thus always destroys sparsity [22].

IV. Compressible Sources: Measuring Non-strict Sparsity

We introduce two ways of *measuring* compressibility (non-strict sparsity) that are both intuitive and useful, in the sense that they will allow us to bound the MSE distortion rate function or the source entropy, and thus to connect the notion of compressibility with actual lossy compression performance.

A. Incomplete Moments as Compressibility Measure

A possible qualitative characterization of compressibility is as follows: for a fixed sample vector $(x_1, x_2, \ldots, x_n)$ of size n, the fewer samples $k \leq n$ are needed to capture a large part of the vector's energy, the more compressible is the vector. This can be quantified by ordering the samples according to their magnitudes, e.g. with a permutation π such that $x_{\pi(i)} \geq x_{\pi(i+1)}$, $1 \leq i < n$, and computing the second moment $\tilde{A}_k/k = \sum_{i=1}^{k} x_{\pi(i)}^2/k$ (the average energy) of the k largest samples. Then a vector with total energy $\tilde{A}_n$ will be more compressible if $\tilde{A}_k/\tilde{A}_n$ grows more rapidly towards 1, in the sense that the distortion of the approximation by the k largest samples, $D_k = \tilde{A}_n - \tilde{A}_k$, will be smaller. For asymptotic block lengths, this approach can be applied to a memoryless continuous source X with density $f(x)$, by considering the proportion of largest samples $\tilde{\mu} \in [0, 1]$ ($\tilde{\mu} = k/n$ for finite lengths) and their second moment $\tilde{A}(\tilde{\mu})/\tilde{\mu}$. A simple parametric way to obtain these quantities is to compute two *incomplete moments* for the realizations above a magnitude threshold t, namely the probability

$$\mu(t) = \int_{-\infty}^{-t} f(x)\,\mathrm{d}x + \int_{t}^{\infty} f(x)\,\mathrm{d}x \tag{7}$$

and the second moment

$$A(t) = \int_{-\infty}^{-t} x^2 f(x)\,\mathrm{d}x + \int_{t}^{\infty} x^2 f(x)\,\mathrm{d}x, \tag{8}$$

where $A(0) = \sigma^2$ is the source variance (we assume $\mathrm{E}\,X = 0$ without loss of generality).

The parametric curve $\mathcal{L}(t) = (\mu(t), A(t)/\sigma^2)$, which runs from $(0, 0)$ to $(1, 1)$ for $t = \infty \ldots 0$, can be used to measure the compressibility of X. The parameter t may be eliminated, yielding the *moment profile* $\tilde{A}(\tilde{\mu}) = A(\mu^{-1}(\tilde{\mu}))$, which is monotonically increasing, concave-$\cap$ for $\tilde{\mu} = 0 \ldots 1$ (see [18, Sec. 4.2], where the moment profile was first proposed to characterize compressible sources). Thus the faster $\tilde{A}$ grows for small $\tilde{\mu}$, the more compressible is the source.

Clearly, characterizing compressibility with a curve instead of a single parameter is a bit cumbersome. One alternative is to determine a special point $(\mu^*, A^*/\sigma^2)$ on $\mathcal{L}$, which yields an upper bound on the differential entropy $h(X)$ (see Corollary 4 in Section V-A). A simpler alternative, which however has no straightforward connection with entropy, is to measure compressibility by the area under the curve $\mathcal{L}$ for $\mu = 0 \ldots 1$. In fact, it turns out that incomplete moments have long been used to measure inequality in distributions, and that $\mathcal{L}$ is basically a Lorenz curve [23] for asymptotically large

samples of the squared random variable X^2. Recent work by Hurley and Rickard [24] compared the *Gini index* – twice the area between $\mathcal{L}$ and the diagonal from $(0,0)$ to $(1,1)$ – with other measures of sparsity and found it to be one of the most useful under a number of criteria.

Section V will present upper bounds on $D(R)$ that can be computed directly from the incomplete moments $\mu(t)$ and $A(t)$, thus relating compressibility with lossy compression.

B. The Geometric Mean as Compressibility Measure

This section introduces the geometric mean, normalized by the standard deviation, as a single-parameter compressibility measure. A sequence of n positive real numbers, $x_1, x_2, \ldots, x_n$, has arithmetic mean $A_n = \frac{1}{n}\sum_{i=1}^n x_i$ and geometric mean $G_n = (\prod_{i=1}^n x_i)^{1/n}$. The classic arithmetic-geometric mean inequality is $G_n \leq A_n$, with equality if and only if all x_i are equal. The geometric mean equals the side length of an n-cube with the same volume as the rectangular parallelepiped spanned by the x_i. A small ratio G_n/A_n corresponds to a "thin" parallelepiped or a sparse, compressible sequence $\{x_i\}$. Conversely, $G_n/A_n = 1$ yields an n-cube, that is the least sparse sequence $\{x_1 = x_2 = \ldots = x_n\}$. We will use the expected geometric mean of a block of sample magnitudes, in the limit of large block length, to measure the compressibility of a memoryless source.

Definition 3 *The geometric mean of a memoryless continuous source X with $\Pr\{X{=}0\} = 0$ is $G(X) = \exp(\mathrm{E}\log|X|)$.*

To see that $G(X)$ is well defined for a memoryless source X with density $f(x)$ and is indeed the desired quantity, consider a block of n i.i.d. samples from X. The geometric mean of these n samples is $G_n(\boldsymbol{x}) = (\prod_{i=1}^n x_i)^{1/n}$, while its expected value is

$$\begin{aligned}\mathrm{E}\, G_n(\boldsymbol{X}) &= \int \prod_{i=1}^n |x_i|^{1/n} \prod_{i=1}^n f(x_i)\,\mathrm{d}\boldsymbol{x} \\ &= \prod_{i=1}^n \int |x_i|^{1/n} f(x_i)\,\mathrm{d}x_i = (\mathrm{E}\,|X|^{1/n})^n,\end{aligned}$$

since Fubini's theorem can be applied to the product density. If we let the block size go to infinity, we obtain the geometric mean of the source [25, p. 139]:

$$\begin{aligned}G(X) &= \lim_{n\to\infty} \left(\mathrm{E}\,|X|^{1/n}\right)^n \\ &= \lim_{p\to 0+} \left(\mathrm{E}\,|X|^p\right)^{1/p} = \exp(\mathrm{E}\log|X|). \end{aligned} \tag{9}$$

For a fixed source variance, if more probability mass is concentrated around zero, $G(X)$ will become smaller and a sample vector of X will look sparser. Due to the fixed variance, the density will become more heavy-tailed at the same time.

Different sparsity (compressibility) measures have been proposed for a variety of applications: a quite common one is the quasi-norm $\|\boldsymbol{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ with $0 < p \leq 1$; see for example [22], [24] and references therein. The obvious question is: how to choose p? If $\boldsymbol{x}$ is a sample from a memoryless source, choosing $p = 1/n$ will yield the geometric mean as $n \to \infty$, by equation (9). This is a strong argument in favor of the geometric mean as a compressibility measure for continuous random variables. In this respect, it is also interesting to observe that for vectors from a bounded set, $\lim_{p\to 0+} \|\boldsymbol{x}\|_p^p$ is equal to the Hamming weight $w_H(\boldsymbol{x})$, which is the strictest sparsity measure in the sense that only values that are exactly zero contribute to sparsity (cf. Donoho's l_0 "norm").

Section VI will show that the geometric mean in combination with the variance can be used to bound the source entropy and therefore characterize asymptotic $R(D)$ behavior.

V. Compressible Sources: Distortion Rate Bounds

A. Two Upper Bounds

This section presents two upper bounds on the MSE $D(R)$ of continuous random variables, which will be applied to models of compressible sources in the following sections. The bounds are obtained by classifying the magnitudes of the source samples using a threshold t and applying the Gaussian upper bound (2) to each of the two classes. They are upper bounds on the operational rate distortion function of *magnitude classifying quantization* (MCQ), which sends the classification as side information and uses it to switch between two codebooks. The samples with magnitude above threshold are called *significant* and are characterized by the two incomplete moments used to measure sparsity in Section IV-A, namely the probability $\mu(t)$ (7) and the second moment $A(t)$ (8), where $A(0) = \sigma^2$ is the source variance. From these we compute the conditional second moment of the significant samples,

$$\sigma_1^2(t) = \mathrm{E}[X^2 \,|\, |X| \geq t] = \frac{A(t)}{\mu(t)},$$

as well as that of the *insignificant* samples,

$$\sigma_0^2(t) = \mathrm{E}[X^2 \,|\, |X| < t] = \frac{\sigma^2 - A(t)}{1-\mu(t)}.$$

The classification decision is sent as side information to the decoder, using $h_b(\mu)$ nats per sample. The encoder can now use two separate Gaussian codebooks, one for the significant samples with rate R_1 and one for the insignificant samples with rate R_0. The average rate per sample becomes

$$R = h_b(\mu(t)) + \mu(t)R_1 + (1-\mu(t))R_0. \tag{10}$$

By standard rate allocation (reverse water-filling) over the two codebooks we obtain an upper bound.

Theorem 3 (High-Rate Upper Bound) *For all*

$$R \geq R_{\min}(t) = h_b(\mu(t)) + \frac{1}{2}\mu(t)\log\frac{\sigma_1^2(t)}{\sigma_0^2(t)}, \tag{11}$$

the MSE distortion rate function of a memoryless continuous source is upper-bounded by

$$D(R) \leq B_{hr}(t,R) = c(t)\sigma^2 e^{-2R}, \tag{12}$$

where

$$c(t) = \exp\left(2h_b(\mu(t)) + (1-\mu(t))\log\tfrac{\sigma_0^2(t)}{\sigma^2} + \mu(t)\log\tfrac{\sigma_1^2(t)}{\sigma^2}\right). \tag{13}$$

The best asymptotic upper bound for $R \to \infty$ is obtained by finding the threshold $t^* \geq 0$ that minimizes $c(t)$. Since $\lim_{t\to 0+} c(t) = 1$, the Gaussian upper bound is always a member of this family.

Proof: The variances of the insignificant and the significant samples can be upper-bounded by the second moments, as $\mathrm{Var}(X||X|<t) \leq \sigma_0^2(t)$ and $\mathrm{Var}(X||X|\geq t) \leq \sigma_1^2(t)$. By inserting these into the Gaussian upper bound (2) and weighting with the respective probabilities, we obtain

$$\mathrm{E}\, d(X,\hat{X}) \leq (\sigma^2 - A(t))e^{-2R_0} + A(t)e^{-2R_1}.$$

For given t, the optimal split of the total rate (10) can be found using Lagrangian optimization. The condition $R \geq R_{\min}$ ensures that the rates R_0 and R_1 are nonnegative. ∎

Exploiting the trivial fact that (12) also upper bounds the Shannon lower bound $D_{\mathrm{SLB}}(R) = e^{2h(X)-2R}/2\pi e$, we obtain an upper bound on differential entropy.

Corollary 4 *Let $\mu^* = \mu(t^*)$ and $A^* = A(t^*)$ yield the tightest bound in Theorem 3. Define the pmf's*

$$\boldsymbol{\mu}^* = [\mu^*, 1-\mu^*], \quad \boldsymbol{a}^* = \left[\tfrac{A^*}{\sigma^2}, 1-\tfrac{A^*}{\sigma^2}\right].$$

Then the differential entropy $h(X)$ is upper-bounded by

$$h(X) \leq \tfrac{1}{2}\ln(2\pi e\sigma^2) + h_b(\mu^*) - \tfrac{1}{2}D(\boldsymbol{\mu}^*\|\boldsymbol{a}^*), \qquad (14)$$

where $D(\cdot\|\cdot)$ is the divergence or Kullback-Leibler distance between the pmf's.

For $t^* = 0$, that is $\mu^* = 1$, the bound (14) reduces to the Gaussian upper bound on entropy. Highly compressible sources with a peaked, heavy-tailed pdf will have a much smaller entropy than a Gaussian with the same variance. In that case the divergence term in (14) will be large, and the side information term $h_b(\mu^*)$ becomes negligible. In a certain sense this entropy bound generalizes and quantifies the concept that the more confined a distribution is, the smaller its entropy [14, Sec. 20].

A low-rate bound is obtained by upper-bounding only the significant samples, while the other samples are quantized to zero, thus yielding a distortion floor.

Theorem 5 (Low-Rate Upper Bound) *The MSE distortion rate function of a memoryless continuous source is upper-bounded by*

$$D(R) \leq B_{lr}(t,R), \quad \text{for } t \geq 0 \text{ and } R \geq 0 \qquad (15)$$

where

$$B_{lr}(t,R) = A(t)\exp\left(-2\tfrac{R-h_b(\mu(t))}{\mu(t)}\right) + \sigma^2 - A(t).$$

For a given threshold $t \geq 0$, satisfying the condition given hereafter, this bound can be optimized to yield

$$D(R^*(t)) \leq B_{lr}(t, R^*(t)), \qquad (16)$$

with the locally optimal rate (with respect to t) given by

$$R^*(t) = h_b(\mu(t)) - \tfrac{1}{2}\mu(t)\Big[2h_b'(\mu(t)) + \gamma(t) + \mathrm{W}_{-1}\left(-\gamma(t)e^{-2h_b'(\mu(t))-\gamma(t)}\right)\Big], \qquad (17)$$

where γ is the reciprocal normalized second tail moment

$$\gamma(t) = \frac{\mu(t)}{A(t)}t^2 = \frac{t^2}{\mathrm{E}[X^2||X|\geq t]} \qquad (18)$$

and W_{-1} is the second real branch of the Lambert W *function, taking values on $(-\infty,-1]$. (The function $\mathrm{W}(x)$ solves $\mathrm{W}(x)e^{\mathrm{W}(x)} = x$.)*

The condition on t in order for $R^*(t)$ to be well-defined is that the argument of W_{-1} in (17) be larger than or equal to $-1/e$, that is, $-\gamma(t)e^{-2h_b'(\mu(t))-\gamma(t)} \geq -1/e$. The rate $R^*(t)$ is only locally optimal in the sense that a small variation of t will not tighten (16)*, but there might exist $t' \neq t$ such that the corresponding bound is strictly tighter at $R = R^*(t)$.*

The proof appears in Appendix C, followed by detailed discussions of when (17) has no solution, i.e. is not well-defined, as well as its locally optimal character. Furthermore, a corollary shows that the low-rate and high-rate bounds coincide in the minimum of the latter, that is, as expected there is a continuous transition between the two bounds. Expression (17) can be simplified, at the price of yielding a looser bound, by replacing W with an approximation [18, Sec. 2.5].

One may use (16) to trace an upper bound on $D(R)$ by sweeping the threshold $t = \infty \ldots t^*$, that is going from $R = 0$ to $R = R_{\min}(t^*)$, at which point the high-rate bound (12) takes over, i.e. is tighter for all $R > R_{\min}(t^*)$. Results by Sakrison [26] and Gish and Pierce [27] imply that the operational distortion rate function $\delta(R)$ of a magnitude classifier followed by a Gaussian scalar quantizer (adapted to the class variance) will be at most a factor of $\pi e/6$ (1.53 dB) above these bounds. Actually, this gap is even smaller at low rates, since the distortion $D_0(0) = \sigma_0^2$ is trivially achieved for the insignificant samples.

The high-rate bound (12) does not apply to the spike source (5), since its distribution is not continuous, cf. the discussion following (5). However, the low-rate bound (16) holds for any threshold $t > 0$, as the significant samples then have a continuous density. In the limit of arbitrarily small positive t, such that $\mu \to p$, (15) becomes the simple spike upper bound (6). For many sparse sources, the low-rate bound (16) turns out to be much tighter than the Gerrish-Schultheiss bound [28]; see [18, Sec. 3.5] for an example.

The bounds can also be computed directly from the moment profile $\tilde{A}(\tilde{\mu})$, without resorting to the underlying source pdf, since $\frac{\mathrm{d}\tilde{A}(\tilde{\mu})}{\mathrm{d}\tilde{\mu}}\big|_{\tilde{\mu}=\mu(t)} = t^2$ is the only additional quantity needed to compute (16) [18, Sec. 4.2]. This reinforces the usefulness of the moment profile as a measure of compressibility.

For most source densities it is very difficult, if not impossible, to compute the distortion rate function in closed form. A popular escape route is to discretize the density and apply the Blahut-Arimoto algorithm to compute a numerical approximation of $D(R)$ [10, Sec. 13.8]. To obtain plausible results, one needs to pay close attention to the discretization

and to the artifacts due to finite entropy (i.e. distortion falsely dropping to zero), particularly for highly compressible sources. Thus the bounds presented here can be a valuable alternative, since the required incomplete moments can be easily computed for most densities, at least numerically. Perhaps even more interesting is the possibility to compute *empirical distortion rate bounds* from a sample of the source, from which the needed quantities t, $\mu(t)$ and $A(t)$ can be easily estimated.

B. Application to a Power-Law Source Model

As an example, we apply the above bounds to a power-law model for wavelet coefficients studied e.g. in [9], [7, Sec. 11.4]. The rate $R^*_{\min} = R_{\min}(t^*)$ in the optimized high-rate bound (12) provides an estimate of the beginning of the high-rate region, in which distortion decays with -6 dB/bit. Together with the corresponding distortion bound $B^*_{hr} = B_{hr}(t^*, R^*_{\min})$ it localizes the end of the typical knee between the low-rate region with fast distortion decay and the high-rate region.

Consider a normalized order statistic $m(z)$ that ranks the magnitudes of wavelet image transform coefficients in decreasing order according to the normalized rank z, such that $m(0)$ is the largest coefficient magnitude and $m(1)$ the smallest. The power-law model is based on the empirical observation that the magnitudes of the larger half ($z = 0 \dots 0.5$) decay approximately like a negative power of z, that is $m(z) \approx Cz^{-\gamma}$ up to about $z = 0.5$. The exponent γ is on the order of 1 for typical images.

The connection with the high-rate bound is made by noticing that the classification threshold t divides the coefficients (now thought to come from a memoryless source) into two groups, one with magnitudes above t and one with magnitudes below t. The expected rank of a coefficient with magnitude t will be $\mu(t)$. Thus we may equate $z = \mu(t)$ and $m(\mu(t)) = t$. As mentioned in Section IV-A, the threshold t can be eliminated altogether by substituting it with μ in the integral defining A, yielding the moment profile $\tilde{A}(\tilde{\mu}) = \int_0^{\tilde{\mu}} m^2(z)\,\mathrm{d}z$. Since $z^{-\gamma}$ is generally not square integrable, we change the model to $m(z) = C(\mu_0+z)^{-\gamma}$, where μ_0 is a positive constant that ensures integrability. Finally, the coefficient decay above $z = 0.5$ is observed to be almost linear (i.e. the magnitudes of the 50% smallest coefficients are almost uniformly distributed). This results in the following composite model for the moment profile:

$$\tilde{A}(\tilde{\mu}) = \begin{cases} \int_0^{\tilde{\mu}} C^2(\mu_0+z)^{-2\gamma}\,\mathrm{d}z, & 0 \le \tilde{\mu} \le 0.5, \\ \int_0^{0.5} C^2(\mu_0+z)^{-2\gamma}\,\mathrm{d}z & \\ \quad + \int_{0.5}^{\tilde{\mu}} 4m^2(0.5)(1-z)^2\,\mathrm{d}z, & 0.5 < \tilde{\mu} \le 1. \end{cases} \tag{19}$$

The median magnitude $m(0.5)$ and the exponent γ can be estimated from a sample, the normalization constant is $C = m(0.5)(\mu_0 + 0.5)^{\gamma}$, while μ_0 can be determined numerically from the condition $\tilde{A}(1) = \sigma^2$.

The distortion rate bounds can be computed based on the moment profile $\tilde{A}(\tilde{\mu})$ alone, but if needed the implicit probabilistic source model can be easily deduced. The pdf $g(m)$ of the magnitudes is obtained parametrically from $\frac{\mathrm{d}\tilde{A}(\tilde{\mu})}{\mathrm{d}\tilde{\mu}} = m^2$ and $\frac{\mathrm{d}^2\tilde{A}(\tilde{\mu})}{\mathrm{d}\tilde{\mu}^2} = -\frac{m}{g(m)}$. If desired, a symmetric source model pdf is $f(x) = \frac{1}{2}g(|x|)$.

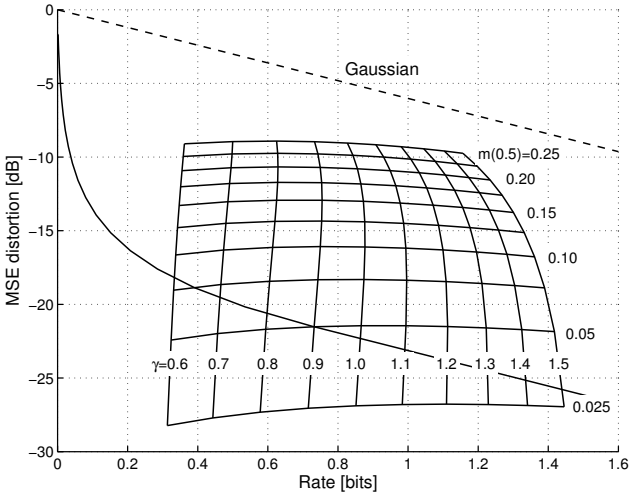


Fig. 4. Beginning of high-rate region, i.e. approximate location where the knee in the $D(R)$ curve ends, for the power-law wavelet coefficient model (19). The points of the grid indexed by the model parameters (exponent γ and median magnitude $m(0.5)$) correspond to the points (R, D) where the optimized high-rate bound (12) starts to hold, i.e. starting from where -6 dB/bit decay is predicted. Also shown are the bound (16) for rates up to $R = 0.72$ bits, and the bound (12) for rates above $R = 0.72$ bits (solid line), for the parameters $\gamma = 0.9$, $m(0.5) = 0.05$, as well as the Gaussian $D(R)$ (all with unit variance).

Fig. 4 displays a grid of points $(R^*_{\min}, B^*_{hr})$ obtained from the model (19) for a range of parameter values. Interestingly, the parameters have nearly orthogonal influences over a wide range: the exponent γ affects mainly the rate $R^*_{\min}$, while the median magnitude $m(0.5)$ affects the distortion B^*_{hr}. In terms of the source pdf, a small $m(0.5)$ implies that most of the source energy is in the pdf tail; in turn, γ controls the tail decay, which will be slower for smaller γ ("heavy tail"). Points that lie on a line with slope -6 dB/bit correspond to asymptotically equal upper bounds, i.e. to sources that can be compressed equally well at high rates. The median $m(0.5)$ can be seen as an indicator of sparsity that has a strong influence on asymptotic compressibility. The exponent γ controls how fast the asymptotic regime is reached; to have high compression at low rates, both $m(0.5)$ and γ need to be small, i.e. the source pdf must be peaked at zero and heavy-tailed at the same time.

Also shown in Fig. 4 are the bounds (12) and (16) for $\gamma = 0.9$, $m(0.5) = 0.05$, which are the approximate parameters of the wavelet coefficients used to draw Fig. 1. The estimated start of the high-rate region is $R^*_{\min} = 0.72$ bits, $B^*_{hr} = -21.5$ dB, matching quite well with Fig. 1. Due to the roundness of the knee it is hard to visually estimate where the asymptotic decay of -6 dB/bit begins. Gaussian mixture models (Section VII) may show much sharper knees.

The power-law model (19) provides a valuable empirical tool for analyzing wavelet coefficients or other approximately scale-invariant data. However, it lacks the generality and versatility, as well as the theoretical apparatus, of the Gaussian mixture models that will be studied in Section VII. In particular, not every sparsifying transform will necessarily produce the approximately scale-invariant coefficients implied by the power-law model.

VI. Compressible Sources: Entropy Bounds

The geometric mean introduced as a measure of compressibility in Section IV-B can be used to obtain bounds on source entropy.

A. Lower Bounds on Differential Entropy

The logarithm of the geometric mean, $\mathrm{E}\log|X|$, yields a lower bound on the entropy of continuous random variables with one- or two-sided monotone densities. Through the SLB (3), this can be used to bound asymptotic $R(D)$ for $D \to 0$.

We first prove a weaker bound that has the appeal of displaying the relationship with an analogous bound for discrete entropy. Then we will prove a bound which is tight for the class of monotone densities considered.

Notice that in general the geometric mean has to be normalized by the standard deviation, $\sigma = \sqrt{\mathrm{Var}(X)}$, before it can be used as sparsity measure. However, in the following results this is not done, since the entropy would also have to be normalized (as in $h(\sigma^{-1}X) = h(X) - \log\sigma$) and the two normalizations cancel each other.

Proposition 6 *Let X be a finite variance random variable with a monotone one-sided pdf f and domain $[x_0, \infty)$ or $(-\infty, x_0]$. Then*

$$h(X) \geq \mathrm{E}\log|X - x_0|.$$

Proof: Without loss of generality, consider a pdf f which is monotone non-increasing on $[x_0, \infty)$. The monotonicity implies that f is Riemann-integrable, and the finite variance ensures that the entropy integral is finite (by the Gaussian upper bound on entropy, $h(X) \leq \frac{1}{2}\log(2\pi e\sigma^2)$, [10, Thm. 9.6.5]). We will approximate the integral $h(X) - \mathrm{E}\log|X - x_0| = -\int_{x_0}^{\infty} f(x)\log(|x - x_0|f(x))\,\mathrm{d}x$ by a Riemann sum with step size Δ. Let $x_i = x_0 + i\cdot\Delta$ and $p_i = f(x_i)\Delta$, for $i = 1, 2, \ldots$. By monotonicity, we have $p_1 \geq p_2 \geq \ldots$ and hence

$$1 \geq \sum_{i=1}^{\infty} p_i \geq \sum_{i=1}^{n} p_i \geq np_n. \tag{20}$$

Thus we can write

$$\begin{aligned} h(X) - \mathrm{E}\log|X - x_0| &= \lim_{\Delta\to 0} -\sum_{n=1}^{\infty} p_n \log(|x_n - x_0|f(x_n)) \\ &= \lim_{\Delta\to 0} -\sum_{n=1}^{\infty} p_n \log\left(n\Delta\cdot\frac{p_n}{\Delta}\right) \\ &\geq \lim_{\Delta\to 0} -\sum_{n=1}^{\infty} p_n \log(1) = 0, \end{aligned}$$

where the inequality follows from taking the logarithm of (20). ∎

Remark: Inequality (20) was used by Wyner to prove an analogous bound for discrete entropy [29].

Using a different proof technique, we obtain a stronger result:

Theorem 7 *Let X be a finite variance random variable with a monotone one-sided pdf f and domain $[x_0, \infty)$ or $(-\infty, x_0]$. Then*

$$h(X) \geq \mathrm{E}\log|X - x_0| + 1, \tag{21}$$

with equality if and only if f is a uniform density.

Proof: For simplicity, we assume f to be non-increasing on $[0, \infty)$. Let $\mathcal{B}$ be the set of all such monotone non-increasing, finite variance densities on $[0, \infty)$. It is easy to verify that $\mathcal{B}$ is a convex set. Its boundary $\partial\mathcal{B}$ is the set of all finite variance uniform densities:

$$\partial\mathcal{B} = \{u(a, x) : a \in (0, \infty)\}, \tag{22}$$

where

$$u(a, x) = \begin{cases} 1/a & \text{if } 0 \leq x \leq a, \\ 0 & \text{else.} \end{cases}$$

To see that (22) is indeed the boundary of $\mathcal{B}$, observe first that no uniform density $u(a, x)$ can be written as a nontrivial convex combination of two distinct monotone non-increasing densities. Moreover, any once differentiable $f \in \mathcal{B}$ can be written as a convex combination of elements of $\partial\mathcal{B}$:

$$f(x) = \int_0^{\infty} \lambda(a)u(a, x)\,\mathrm{d}a, \tag{23}$$

where $\lambda(a) = -af'(a)$, as can be shown with some simple calculus. λ is a proper density if f has finite variance (in particular, $\lim_{x\to\infty} xf(x) = 0$) and if $f'(x) \leq 0$, which is indeed the case for monotone decreasing f. Using the standard extensions to distributions, (23) also holds if f contains a countable number of steps, e.g. if it is piecewise constant. In fact, (23) is nothing but a disguised version of the "layer cake" representation[1] of f, namely $f(x) = \int_0^{\infty} \chi_{\{f>t\}}(x)\,\mathrm{d}t$, where $\chi_{\{f>t\}}(x)$ is the indicator function of the level set $\{f(x) > t\}$. The existence of this representation follows from the monotonicity of f.

Looking at (21), we see that

$$h(X) - \mathrm{E}\log X = -\int_0^{\infty} f(x)\log(xf(x))\,\mathrm{d}x \tag{24}$$

is a concave-$\cap$ functional of f, since $h(X)$ is concave and $\mathrm{E}\log X$ is linear in f. Therefore a minimum of (24) over the convex set $\mathcal{B}$ must necessarily lie on its boundary $\partial\mathcal{B}$. We insert an arbitrary boundary element $u(a, x)$ $(0 < a < \infty)$ in (24) to obtain

$$\begin{aligned} h(X) - \mathrm{E}\log X &= -\int_0^{\infty} u(a, x)\log(xu(a, x))\,\mathrm{d}x \\ &= -\int_0^{a} \tfrac{1}{a}\log\tfrac{x}{a}\,\mathrm{d}x \\ &= \log a - \tfrac{x}{a}(\log x - 1)\big|_0^a \\ &= 1. \end{aligned} \tag{25}$$

Since (25) holds for any a, we conclude that it is the global minimum, thus proving (21) and one part of the "if and only

[1] The term "layer cake" representation stems from the picture of cutting the area between $f(x)$ and the abscissa into thin horizontal stripes with widths corresponding to the level sets [30, Sec. 1.13].

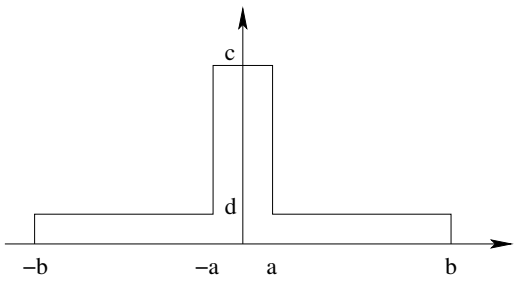


Fig. 5. Probability density (28) of a "uniform spike."

if". To prove the other part, it suffices to observe that $h(X) - \mathrm{E}\log X$ is a strictly concave functional and thus will be larger than (25) in the interior $\mathcal{B} \setminus \partial\mathcal{B}$. ∎

Remark: The tightened bound (21) can be used in turn to tighten Wyner's discrete entropy bound [29], leading to improved performance bounds for a class of lossless codes [31].

Definition 4 *A* weakly unimodal *density with mode x_0 is a pdf which is monotone non-decreasing on $(-\infty, x_0]$ and monotone non-increasing on $[x_0, \infty)$.*

Corollary 8 *Let X be a finite variance random variable with weakly unimodal pdf f such that $\Pr\{X \le x_0\} = \alpha$, where x_0 is the mode. Then*

$$h(X) \ge \mathrm{E}\log|X - x_0| + 1 + h_b(\alpha). \tag{26}$$

For a density that is symmetric about x_0, $f(-x-x_0) = f(x-x_0)$, (26) reduces to

$$h(X) \ge \mathrm{E}\log|X - x_0| + 1 + \log 2. \tag{27}$$

The bound (27) is asymptotically attained by a "uniform spike" with finite variance σ^2 and parameters $0 < a < \sqrt{3}\sigma < b$ defining the density

$$f(x) = \begin{cases} \frac{b(b+a)+a^2-3\sigma^2}{2ab(b+a)} =: c, & |x| \le a, \\ \frac{3\sigma^2-a^2}{2b(b^2-a^2)} =: d, & a < |x| \le b, \\ 0, & \textit{else}, \end{cases} \tag{28}$$

as the tail width $b \to \infty$.

Proof: We view the weakly unimodal pdf f as a mixture of two non-overlapping monotone one-sided densities, $f_l(x)$ and $f_r(x)$, with weights α and $1-\alpha$, respectively. Without loss of generality we can assume $x_0 = 0$. Then,

$$\begin{aligned} h(X) - \mathrm{E}\log|X| &= -\mathrm{E}_f \log[|X| f(X)] \\ &= -\int_{-\infty}^{0} \alpha f_l(x) \log(-x\alpha f_l(x)) \\ &\quad - \int_{0}^{\infty} (1-\alpha) f_r(x) \log(x(1-\alpha) f_r(x)) \\ &= h_b(\alpha) - \alpha\,\mathrm{E}_{f_l} \log[|X| f_l(X)] \\ &\quad - (1-\alpha)\,\mathrm{E}_{f_r} \log[|X| f_r(X)] \\ &\ge h_b(\alpha) + 1, \end{aligned}$$

where the last inequality follows from Theorem 7, proving (26).

It is easily verified that "uniform spikes" exist for $0 < a < \sqrt{3}\sigma < b$, see Fig. 5. Using c and d defined in (28), the asymptotic entropy is

$$\lim_{b\to\infty} h(X) = \lim_{b\to\infty} -2ac\log c - 2(b-a)d\log d = \log(2a)$$

and the asymptotic logarithm of the geometric mean is

$$\begin{aligned} \lim_{b\to\infty} \mathrm{E}\log|X| &= \lim_{b\to\infty} \left[2c\int_0^a \log x\,\mathrm{d}x + 2d\int_a^b \log x\,\mathrm{d}x \right] \\ &= \lim_{b\to\infty} 2ac(\log a - 1) \\ &\qquad + 2d(a - a\log a - b + b\log b) \\ &= \log a - 1. \end{aligned}$$

Hence the lower bound (27) is asymptotically attained by a random variable concentrating its probability uniformly over $[-a, a]$ (since $\lim_{b\to\infty} 2ac = 1$), with an infinite tail contributing only to its variance. A peakier density with the same variance and entropy will have a smaller geometric mean. ∎

Remark: Since only monotonicity and finite variance are needed for Theorem 7 to hold, it can be seen that Corollary 8 holds also for bounded random variables with range $[x_{\min}, x_{\max}]$ and a pdf f that is monotone non-increasing on $[x_{\min}, x_0]$ and monotone non-decreasing on $[x_0, x_{\max}]$ (e.g. a "bathtub" shape).

B. Upper Bound on Differential Entropy

If both the variance and the geometric mean are known, an upper bound on the entropy can be easily obtained via the maximum entropy approach. Owing to the assumptions made in this variational approach, the results in this subsection hold for random variables which have an absolutely continuous distribution function $F(x)$ with probability density $f(x) = F'(x)$.

Proposition 9 *The maximum entropy pdf given the constraints $\mathrm{E}\,X^2 = \sigma^2$ and $\mathrm{E}\log|X| = \theta$ is*

$$f(x) = [\Gamma(\tfrac{u}{2})]^{-1} \left(\tfrac{u}{2\sigma^2}\right)^{u/2} |x|^{u-1} \exp\left(-\tfrac{ux^2}{2\sigma^2}\right), \tag{29}$$

where $\Gamma(z)$ is the gamma function (Euler's integral of the second kind) defined as $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\,\mathrm{d}t$ ($\mathrm{Re}\ z > 0$) [32, 8.31]. The shape parameter $u > 0$ is obtained by solving

$$\mathrm{E}\log|X| = \tfrac{1}{2}\psi(\tfrac{u}{2}) - \tfrac{1}{2}\log\tfrac{u}{2\sigma^2} = \theta, \tag{30}$$

where $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$ [32, 8.36]. For any $\theta \le \log\sigma$ there is a unique solution, since $\mathrm{E}\log|X|$ is strictly monotone increasing in u. The resulting entropy is

$$h(\sigma, \theta) = \tfrac{u}{2} - \tfrac{u-1}{2}\psi(\tfrac{u}{2}) + \log\Gamma(\tfrac{u}{2}) - \tfrac{1}{2}\log\tfrac{u}{2\sigma^2}. \tag{31}$$

Setting $u = 1$ yields the Gaussian density and thus the global entropy maximum given the variance constraint alone.

The proof appears in Appendix D.

Corollary 10 *The entropy of any random variable X with probability density f satisfying $\mathrm{E}\,X^2 = \sigma^2$ and $\mathrm{E}\log|X| = \theta$ is upper bounded by (31).*

The corollary is implied by the maximum entropy approach.

Theorem 11 *The maximum entropy (31) for a finite variance σ^2 has the following asymptotic behavior as the geometric mean e^θ goes to zero, resp. $\theta \to -\infty$:*

$$h(\sigma,\theta) \sim \theta + \log(-2e\theta) \quad \text{as } \theta \to -\infty.$$

The symbol $\sim$ denotes asymptotic equality, i.e. $f \sim g$ as $x \to \infty$ means that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$.

Proof: Note that $\theta \to -\infty$ corresponds to $u \to 0+$. Let

$$\begin{aligned}\Delta &= h(\sigma,\theta) - \theta - \log(-2e\theta) \\ &= \frac{u}{2} - \frac{u}{2}\psi\left(\frac{u}{2}\right) - 1 + \log\left(\frac{\Gamma(\frac{u}{2})}{-\psi(\frac{u}{2}) + \log\frac{u}{2\sigma^2}}\right). \quad (32)\end{aligned}$$

To prove $\lim_{u\to 0+}\Delta = 0$, which is slightly stronger than required, we use the functional relationships $\Gamma(x+1) = x\Gamma(x)$, $\psi(x+1) = \psi(x) + \frac{1}{x}$ and the truncated series expansions $\Gamma(x+1) = 1 - \gamma x + O(x^2)$, $\psi(x+1) = -\gamma + \frac{\pi^2}{6}x + O(x^2)$, both for $|x| < 1$ (see e.g. [32, 8.3]; $\gamma = 0.5772\ldots$ is Euler's constant). We have

$$\lim_{u\to 0+} \tfrac{u}{2}\psi(\tfrac{u}{2}) = \lim_{u\to 0+}[-1 - \gamma\tfrac{u}{2} + \tfrac{\pi^2}{24}u^2 + O(u^3)] = -1,$$

hence $\lim_{u\to 0+}\Delta$ is equal to the limit of the logarithm in (32). But

$$\lim_{u\to 0+} \frac{\Gamma(\frac{u}{2})}{-\psi(\frac{u}{2}) + \log\frac{u}{2\sigma^2}} = \lim_{u\to 0+} \frac{\frac{2}{u}(1 - \frac{\gamma}{2}u + O(u^2))}{\frac{2}{u} + \log\frac{u}{2\sigma^2} + \gamma - \frac{\pi^2}{12}u + O(u^2)} = 1.$$

This can be easily seen by extending the fraction by $\frac{u}{2}$ and observing that $\lim_{u\to 0+} u\log u = 0$. By putting these steps together we obtain $\lim_{u\to 0+}\Delta = 0$. ∎

Fig. 6 shows the lower bound (27) and the upper bound (31) as a function of $\theta = \mathrm{E}\log|X|$ for unit-variance random variables with symmetric unimodal densities. The global maximum of the upper bound corresponds to the unit-variance Gaussian density, which has $\theta \approx -0.635$. As a consequence of Theorem 11, the gap between the lower and upper bounds is asymptotically equal to $\log(-\theta)$. The crossing between upper and lower bounds is only a seeming contradiction, because in fact it simply means that to the right of the crossing there exist no unimodal densities satisfying both the geometric mean and variance constraints. Also shown are the points $(\theta, h(X))$ corresponding to the family of unit-variance generalized Gaussian pdf's $f(t) = \beta/(2\alpha\Gamma(\beta^{-1}))\exp\left(-(|t|/\alpha)^\beta\right)$ with β as a parameter. It can be shown that for $\beta \to 0+$ one has $\theta = \mathrm{E}\log|X| \to -\infty$ and $h(X)$ lies asymptotically halfway between upper and lower bound at distance $\frac{1}{2}\log(-\theta)$.

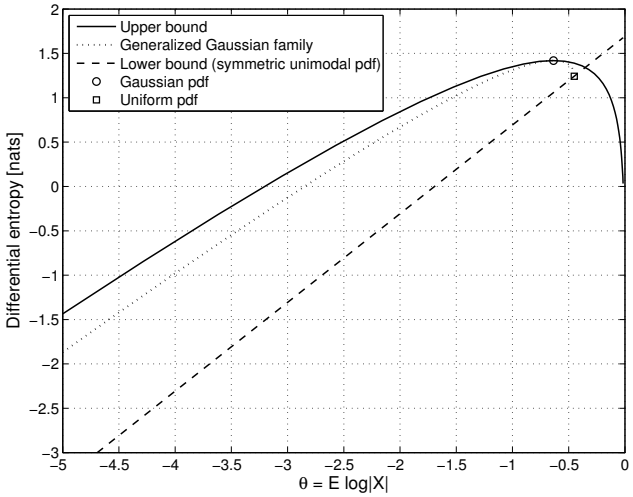


Fig. 6. Differential entropy bounds for symmetric weakly unimodal densities (normalized to unit variance). The square denotes the uniform density for which the lower bound is tight.

VII. Compressible Sources: Gaussian Mixture Models and Coding Gain

The discussion on spikes in Section III-B pointed out that continuous densities are more appropriate for modeling compressible transform coefficients. Gaussian mixtures are a popular approach to model and estimate unknown densities and have been used quite successfully in various applications, see e.g. [33] and references therein. In this section we will study a simple memoryless Gaussian mixture (GM) source model with pdf

$$f(x) = \sum_{s=1}^{N} w_s f_s(x), \quad (33)$$

mixing N zero-mean Gaussian components with variances $\sigma_{m,s}^2$,

$$f_s(x) = \frac{1}{\sqrt{2\pi}\sigma_{m,s}} e^{-x^2/2\sigma_{m,s}^2},$$

with weights $w_s \geq 0$ satisfying $\sum_{s=1}^N w_s = 1$. The spike model of Section III-B may be regarded as a special case of a two-component GM, where one source has zero variance.

For a general GM source X (with possibly nonzero component means), the Shannon lower bound (3) is tight for all $D < D^* = \min_s\{\sigma_{m,s}^2\}$, since then X may be expressed via a "backward test channel" as the sum of a GM with variances $\{\sigma_{m,s}^2 - D\}$ and independent noise $Z \sim \mathcal{N}(0, D)$ [28]; D^* is also known as *critical distortion*. Thus the asymptotic $D(R)$ behavior is determined by the GM entropy, which in general cannot be expressed in closed form. This motivates the bounds presented in the first two subsections, which are followed by a discussion of the relationship with coding gain and some examples.

A. Distortion Rate Bounds for Gaussian Mixtures

The upper bounds introduced in Section V are easily computed for GM models, but they do not exploit the particular

model structure. A GM source may be viewed as containing a hidden discrete memoryless source S that switches between $|\mathcal{S}| = N$ Gaussian sources $\mathcal{N}(0, \sigma_{m,s}^2)$ with selection probabilities $w_s = Pr\{S = s\}$. A lower bound on $D(R)$ is found by assuming that an oracle provides the hidden variable S to the source encoder. Since $S \to X \to \hat{X}$ form a Markov chain, we have

$$I(X; \hat{X}|S) \leq I(X; \hat{X}),$$

where the conditional mutual information is defined as $I(X; \hat{X}|S) = \mathrm{E} \log \frac{f(X,\hat{X}|S)}{f(X|S)f(\hat{X}|S)}$. Computing the lower bound $R_{lb}(D) = \min_{p(\hat{x}|x,s)\in\mathcal{Q}_D} I(X; \hat{X}|S)$, with $\mathcal{Q}_D = \{p(\hat{x}|x, s) : \mathrm{E}(X - \hat{X})^2 \leq D\}$, is equivalent to solving the following standard rate allocation problem:

$$D_{lb}(R) = \min_{\{R_s\}} \sum w_s \sigma_{m,s}^2 2^{-2R_s} \tag{34}$$

subject to

$$\sum w_s R_s = R \text{ and } R_s \geq 0.$$

This yields the lower bound $D(R) \geq D_{lb}(R)$, which can also be seen as a special case of a conditional rate distortion function [34]. The lower bound may be turned into an upper bound by expanding $I(S, X; \hat{X})$ as follows:

$$\begin{aligned} I(S, X; \hat{X}) &= I(X; \hat{X}) + I(S; \hat{X}|X) = I(X; \hat{X}) \\ &= I(S; \hat{X}) + I(X; \hat{X}|S) \leq I(X; \hat{X}|S) + H(S), \end{aligned}$$

using the fact that the mixing variable S is discrete. Thus we have

$$\min_{\mathcal{Q}_D} I(X; \hat{X}|S) \leq R(D) \leq \min_{\mathcal{Q}_D} I(X; \hat{X}|S) + H(S). \tag{35}$$

Clearly, these bounds are not very tight in the case of a GM with large $N = |\mathcal{S}|$ and close to uniform distribution of S. Using Fano's inequality we may see that if S can be estimated from $\hat{X}$ with low probability of error, then $R(D)$ will be close to the upper bound. Conversely, for large $H(S|X) \leq H(S|\hat{X})$ it will be harder to estimate S and $R(D)$ will be closer to the lower bound.

As an example, we used the EM algorithm to estimate the parameters of a two-component GM (33) modeling the wavelet coefficients of the Lena image transformed with the classic 9/7 biorthogonal wavelet. The parameters obtained are $w_1 = 0.9141$, $\sigma_{m,1}^2 = 0.01207$ and $\sigma_{m,2}^2 = 11.51$ (normalized to unit variance). Plots of bound (16) for $R < R_{\min}(t^*) = 0.82$ bits, bound (12) for $R \geq R_{\min}(t^*)$ and the bounds (35) appear in Fig. 7 together with a numerical estimate of $D(R)$ computed with the Blahut-Arimoto algorithm. The gap in (35) is $H(S) = h_b(w_1) = 0.42$ bits wide. Also shown in Fig. 7 are the (R, D) points achieved by a simple embedded (successive refinement) scalar quantizer (see e.g. [7, Sec. 11.5]), applied to $3 \cdot 10^5$ pseudo-random samples. The significance maps were entropy coded, sign and refinements bits were left uncoded. It can be seen that at low rates, thresholding with simple scalar quantization performs very close to the $D(R)$ optimum.

Up to the typical knee, distortion decays faster than -6 db/bit, since mainly the sparse coefficients from the high-variance source are retained by the thresholding operation.

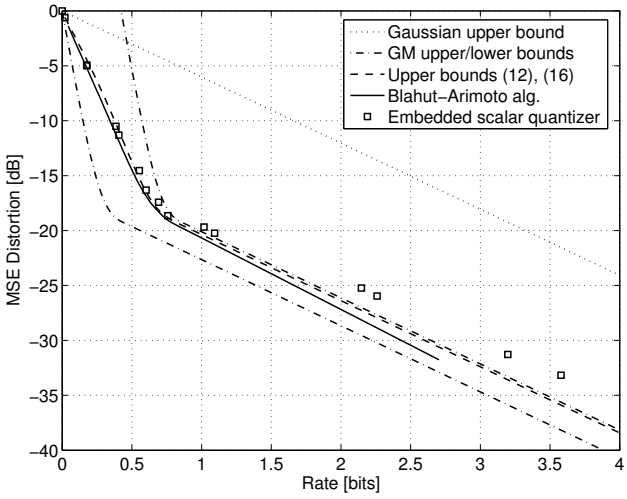


Fig. 7. Distortion rate bounds for two-component Gaussian mixture model of wavelet coefficients.

At higher rates, the coefficients from the low-variance source also start being significant. If the model (33) is extended to $N \geq 3$ Gaussian components, the knee in $D(R)$ becomes rounder, but the basic behavior is unchanged (compare also with Fig. 1 (a)). From these observations we can reach two conclusions: first, two-component GMs suffice to capture the essential features of image coding $D(R)$, and second, the rate $R_{\min}(t^*)$ in the high-rate bound (Theorem 3) is confirmed as estimate of the beginning of the high-rate compression region (see Section V-B). The first observation is also supported by [35], which considers the joint numerical optimization of a classifier and (high-rate) uniform quantizers for each of N classes corresponding to GM components. Simulation results in [35] suggest that for typical image data $N = 2$ components yield a substantial improvement over a single Gaussian, while adding more components gives only minor additional gains.

A general theoretical framework for classified vector quantization (CVQ) of Gaussian mixtures has been introduced by Gray in [36]. Ideally, CVQ would be applied to samples from a multivariate Gaussian mixture having distinct modes, leading to reliable classification. The high-rate bound (12) could be seen as a special case of CVQ, where two mixture components differ only in variance. However, the CVQ approach is different in spirit, since CVQ will usually be applied directly to the data, without first transforming it, while our work focuses on a transform coding setting where a transform generates a sparse signal representation, which is then quantized. The transform coefficients are thought as coming from a sparse memoryless source, which in the above example is shown to be modeled quite well by a mixture of two univariate Gaussian densities.

B. Entropy Bounds for Gaussian Mixtures

The sparsity of a GM source may be measured by the geometric mean, as proposed in Section IV-B, leading to entropy bounds that characterize asymptotic $R(D)$ behavior. The logarithm of the geometric mean of a GM with density (33) is

$$\theta = \mathrm{E}\, log|X| = \tfrac{1}{2}\sum_{s=1}^{N} w_s \log \sigma_{m,s}^2 - \tfrac{1}{2}\log 2 - \tfrac{1}{2}\gamma, \qquad (36)$$

where $\gamma = 0.5772\ldots$ is Euler's constant. The result follows directly from integral 4.333 in [32] and leads to a lower bound on the mixture entropy $h(X)$ via Corollary 8. However, this can be tightened by the same approach as in Section VII-A, namely by lower-bounding the GM entropy by conditioning on the hidden mixing variable:

$$h(X) \geq h(X|S) = \tfrac{1}{2}\sum w_s \log(2\pi e \sigma_{m,s}^2). \qquad (37)$$

This improves the lower bound (27) by the constant $\frac{1}{2}\gamma + \frac{1}{2}\log\frac{\pi}{e}$, as can be seen by inserting (36) into (27) and comparing with (37). From the expansion $I(X;S) = h(X) - h(X|S) = H(S) - H(S|X)$, we obtain the upper bounds

$$h(X) \leq h(X|S) + H(S) \leq h(X|S) + \log N, \qquad (38)$$

with $h(X|S)$ given in (37).

Fig. 8 plots the different bounds that hold for mixtures of zero-mean Gaussians in general and two-component GM in particular, all normalized to unit variance. Also shown is a set of points $(\mathrm{E}\log|X|, h(X))$ corresponding to different two-component GMs. The geometric mean is mainly affected by the ratio $\sigma_{m,2}/\sigma_{m,1}$, while the mixing weights determine $H(S)$ and thus the gap between the lower bound (37) and the tighter upper bound in (38). For large $\sigma_{m,2}/\sigma_{m,1}$, it is easy to estimate S from X and so $h(X)$ will be close to the upper bound. The lower bound can be asymptotically attained with $w_1 \gg w_2$ (then $H(S) \to 0$) for any $\sigma_{m,2}/\sigma_{m,1} \geq 1$; this parallels the "uniform spike" attaining the lower bound in Corollary 8.

Mixtures of a finite number of zero-mean Gaussian components may be considered as a special case of continuous Gaussian scale mixtures [37], which have also been proposed in the context of wavelet coefficient models [1, Sec. VIII.A]. It turns out that the maximum entropy pdf (29) can be expressed as a Gaussian scale mixture.

Proposition 12 *The maximum entropy pdf (29) that satisfies the constraints* $\mathrm{E}\,X^2 = \sigma^2$ *and* $\mathrm{E}\log|X| = \theta$ *can be expressed as a continuous Gaussian mixture*

$$f(x) = \int_0^{\sigma^2/u} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{x^2}{2s^2}} g(s^2)\,\mathrm{d}s^2 \qquad (39)$$

with mixing density

$$g(s^2) = \frac{\sqrt{\pi s^2}(\frac{u}{\sigma^2})^{u/2}}{s^4 \Gamma(\frac{u}{2})\Gamma(\frac{1-u}{2})} \left(\frac{1}{s^2} - \frac{u}{\sigma^2}\right)^{-(1+u)/2}, \qquad (40)$$

where the shape parameter $0 < u < 1$ *is obtained by solving (30).*

Proof: It is easily verified that (29) satisfies the necessary and sufficient conditions for the existence of the representation (39) given in [37] and thus the inversion formula from the same work can be applied, leading to (40). A direct proof can be obtained by substituting $v = 1/\sqrt{s^2}$ in (39) and solving it using integral 3.382(2) in [32]. ∎

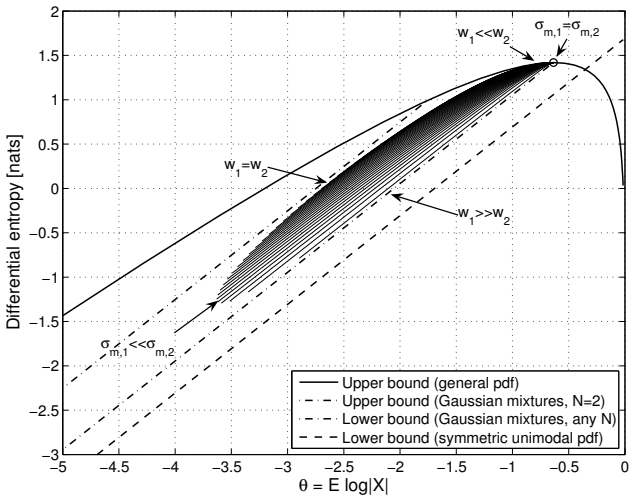


Fig. 8. Differential entropy bounds for two-component Gaussian mixtures, normalized to unit variance. The "cloud" sweeps the pairs $(\mathrm{E}\log|X|, h(X))$ corresponding to $w_1 = 0.01\ldots0.99$ and $\sigma_{m,2}/\sigma_{m,1} = 1\ldots100$. A line in the cloud corresponds to sweeping the variance ratio while keeping the weights fixed.

It is quite surprising that only Gaussian pdf's up to the maximum variance σ^2/u need to be mixed to obtain (29). This is a direct result of the inversion formula, which involves an inverse Laplace transform towards the "time" variable $v = 1/\sqrt{s^2}$, which runs from $\sqrt{u}/\sigma$ to ∞. The mixing density (40) has a "bathtub" shape that concentrates most probability mass close to $s^2 = 0$ and $s^2 = \sigma^2/u$. For $u \to 1$, all mass is shifted towards $s^2 = \sigma^2/u$ (the limit for $u = 1$ is a Gaussian pdf, see Proposition 9), while for $u \to 0$ (very sparse sources) the probability mass is shifted towards $s^2 = 0$. The Laplacian is an example of a pdf that can be represented as a mixture needing component variances going to infinity [37].

C. Coding Gain Revisited

In linear transform coding, the *coding gain* measures the ratio of the asymptotic distortion of a single scalar quantizer to the distortion of a set of quantizers with rate allocation matched to the transform statistics,[2] with both systems operating at the same average rate [39, Sec. 8.7], [38]. Here we will show how the high-rate magnitude classifying quantization (MCQ) upper bound (Theorem 3) leads to an expression that is reminiscent of the coding gain of a transform coding system.

Let us briefly review the derivation of classical transform coding gain. Consider a jointly Gaussian source, emitting independent zero-mean Gaussian random vectors $\boldsymbol{X} = [X_1, X_2, \ldots, X_N]$, with autocorrelation matrix $R_{\boldsymbol{X}}$ such that $R_{ii} = \sigma^2$. (Except for independence, which is not necessary for the derivation, this may be obtained by taking blocks of N samples from a zero-mean weakly stationary discrete

[2] A different definition considers bit allocation for both the original and transformed data [38]. The distinction is significative if the source is non-stationary, e.g. emitting Gaussian random vectors with non-constant autocorrelation vector.

time Gaussian process.) The vector $\boldsymbol{X}$ is multiplied with an orthonormal matrix T, yielding the transformed source vector $\boldsymbol{Y} = T\boldsymbol{X}$, which is then quantized to $\hat{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{W}$, where $\boldsymbol{W}$ models the quantization noise. The reconstructed vector is $\hat{\boldsymbol{X}} = T^{-1}\hat{\boldsymbol{Y}}$. By the Parseval-Plancherel energy conservation formula [7, Sec. A.3], the quantization error in the signal domain will be equal to the error in the transform domain:

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 = \|\boldsymbol{W}\|^2.$$

Also, the average variance of the transform coefficients Y_i is equal to the variance of X:

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\, Y_i^2 = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\, X_i^2 = \sigma^2.$$

This holds (by linearity of expectation) assuming zero mean and can be easily extended to the general case. Let $\sigma_i^2 = \mathrm{E}\, Y_i^2$ be the variance of the i-th transform coefficient. If we use N scalar quantizers to quantize $\boldsymbol{Y}$, the optimal high-rate bit allocation is easily found using Lagrangian optimization.[3] We get an average distortion of the form $D = C\left[\prod_{i=1}^{N}\sigma_i^2\right]^{1/N} e^{-2R}$, with C a constant [38]. This can be compared with the distortion of a scalar quantizer applied to the X_i's, which is $D = C\sigma^2 e^{-2R} = C\left[\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\, X_i^2\right] e^{-2R}$. The *transform coding gain* is now defined as the ratio of the distortion of direct scalar quantization of the source over scalar quantization of the transform coefficients (with bit allocation):

$$\Gamma_{\mathrm{TC}} = \frac{\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2}{\left(\prod_{i=1}^{N}\sigma_i^2\right)^{1/N}} = \frac{A_N(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)}{G_N(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)}. \qquad (41)$$

In purely algebraic terms, equation (41) is the ratio of the arithmetic mean A_N of the coefficient variances to their geometric mean G_N, which might be used as the "axiomatic" definition of coding gain. This short derivation pointed out the implied assumptions, namely high rate and (near-)Gaussianity. In the jointly Gaussian case, using a KLT will maximize the coding gain, that is, minimize the geometric mean of the coefficient variances [7, Sec. 11.3.2].

From the above, it is straightforward to define a measure of MCQ coding gain by considering the ratio of the Gaussian upper bound to the high-rate upper bound (12).

Definition 5 *The coding gain for high-rate optimal*[4] *magnitude classifying quantization is*

$$\Gamma_{\mathrm{MCQ}} = \frac{c(0)}{c(t^*)} = \frac{\sigma^2}{c(t^*)\sigma^2} = \frac{\mu^*\,[\sigma_1^*]^2 + (1-\mu^*)\,[\sigma_0^*]^2}{e^{2h_b(\mu^*)}\,[\sigma_1^*]^{2\mu^*}\,[\sigma_0^*]^{2(1-\mu^*)}}, \qquad (42)$$

where $c(t)$ is as defined in Theorem 3, t^* is the threshold yielding the tightest upper bound and $\mu^* = \mu(t^*)$, $\sigma_0^* = \sigma_0(t^*)$, $\sigma_1^* = \sigma_1(t^*)$.

[3]This relies on the assumption that either the source is jointly Gaussian (then any orthonormal transform will yield Gaussian coefficients), or at least that the signal components X_i and the transform coefficients Y_i have the same marginal high-rate $D(R)$ behavior of the form $D_i = C_i e^{-2R}$. For details on high-rate bit allocation, see e.g. [38] and references therein.

[4]Here *optimal* refers to the tightest upper bound of Theorem 3; directly optimizing a MCQ would yield tighter bounds, because significant and insignificant samples differ in $D(R)$ behavior.

Except for the additional side information factor $e^{2h_b(\mu^*)}$, this definition corresponds to the classical coding gain (41) for two sources with weights μ^* and $1-\mu^*$. This similarity opens a new perspective on transform coding: instead of considering each transform coefficient as a distinct random variable, we mix all coefficients together and use a quantizer for the marginal density. A transform that has high classical coding gain will have a peaked marginal density, so that the MCQ coding gain will also be large. At the same time, the mixing approach obviously entails a loss in coding gain, which we will study by means of an example in Section VII-D.

In general, MCQ will be suboptimal; the following definition allows comparing it with an optimal quantizer that asymptotically achieves the Shannon lower bound.

Definition 6 *The coding gain for a memoryless continuous source X is defined as the ratio of the Gaussian upper bound on $D(R)$ to the Shannon lower bound:*

$$\Gamma_{\mathrm{SLB}} = \frac{2\pi e\sigma^2}{\exp(2h(X))}. \qquad (43)$$

It measures the coding gain achieved by using a codebook matched to the source instead of a Gaussian codebook.

These coding gain definitions are connected with the geometric mean $G(X)$ of a source (Definition 3) through the following.

Definition 7 *The normalized squared geometric mean of a zero-mean memoryless continuous source X is*

$$M_G(X) = \frac{\exp(\mathrm{E}\log X^2)}{\mathrm{E}\, X^2} = \frac{G(X^2)}{A(X^2)} = \frac{[G(X)]^2}{A(X^2)},$$

with the implicit definition $A(X^2) = \lim_{n\to\infty} A_n(X_1^2, X_2^2, \ldots, X_n^2) = \mathrm{E}\, X^2 = \sigma^2$.

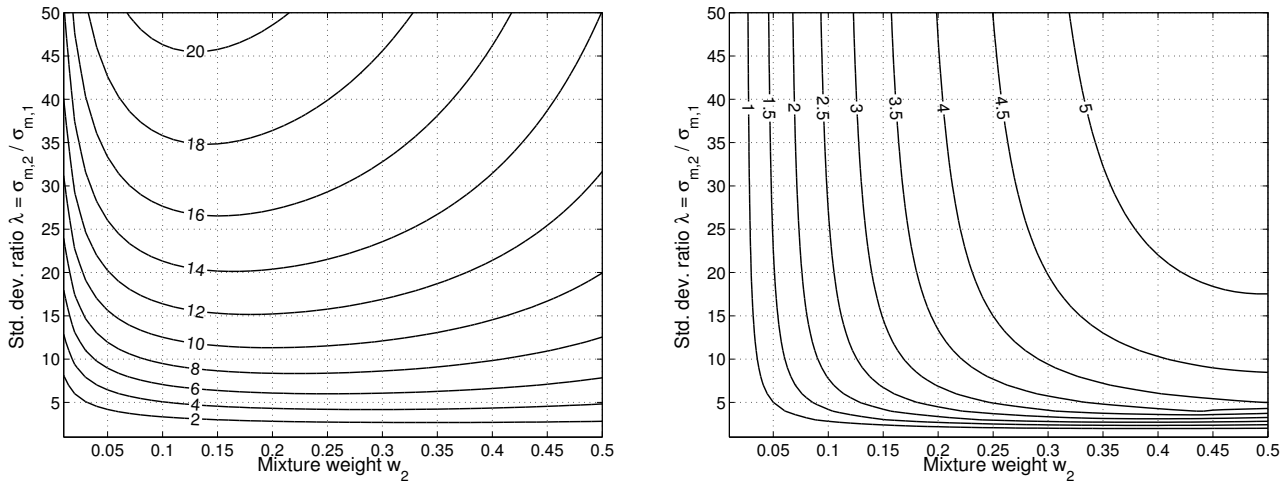
By the arithmetic-geometric mean inequality, $M_G \le 1$, with equality if and only if the source magnitude is constant ($|X| = \sigma$).

Corollary 13 (to Theorem 3) *The factor $c(t)$ in the MCQ high-rate bound (12) is lower-bounded by*

$$c(t) \ge \alpha M_G(X) \quad \textit{for all } t \ge 0,$$

where $\alpha = 1$ for general sources X, $\alpha = \frac{2e}{\pi}$ if X is symmetric weakly unimodal (Definition 4), and $\alpha = 2e^{\gamma}$ if X is a Gaussian mixture of the form (33)*.*

Proof: For Gaussian mixtures, using (36) and (38) we obtain $\frac{\Gamma_{\mathrm{SLB}}}{\Gamma_{\mathrm{MCQ}}} \le \frac{c(t^*)}{2e^{\gamma}M_G}$. By definition we have $\frac{\Gamma_{\mathrm{SLB}}}{\Gamma_{\mathrm{MCQ}}} \ge 1$, so $\alpha = 2e^{\gamma}$ follows. (For $X \sim \mathcal{N}(0,\sigma^2)$, the bound is trivially tight for $t^* = 0$.) For symmetric weakly unimodal X, the geometric mean yields an upper bound on Γ_{SLB} through Corollary 8, leading to $\frac{\Gamma_{\mathrm{SLB}}}{\Gamma_{\mathrm{MCQ}}} \le \frac{c(t^*)\pi}{2eM_G}$ and thus $\alpha = \frac{2e}{\pi}$. Finally, for general X, we bound $\mathrm{E}\log X^2$ by



(a) (b)

Fig. 9. Magnitude classifying quantization (MCQ) of two-component Gaussian mixtures (GM) with component variances $\sigma^2_{m,1}$, $\sigma^2_{m,2}$, and weights $1-w_2$, w_2, respectively. (a) Contours of constant coding gain $\Gamma_{\rm TC}$ (41) (in dB) for *unmixed*, separate sources (equivalent to GM lower bound). (b) Coding gain loss Δ_{CG} (44) (in dB) relative to $\Gamma_{\rm TC}$ for MCQ of the mixture.

applying Jensen's inequality to the significant and insignificant samples separately:

$$\begin{aligned}\mathrm{E}\log X^2 &= \Pr\{X^2<t^2\}\,\mathrm{E}[\log X^2|X^2<t^2] \\ &\quad + \Pr\{X^2\ge t^2\}\,\mathrm{E}[\log X^2|X^2\ge t^2] \\ &\le [1-\mu(t)]\log \mathrm{E}[X^2|X^2<t^2] \\ &\quad + \mu(t)\log \mathrm{E}[X^2|X^2\ge t^2] \\ &= [1-\mu(t)]\log\tfrac{1-A(t)/\sigma^2}{1-\mu(t)} \\ &\quad + \mu(t)\log\tfrac{A(t)/\sigma^2}{\mu(t)} + \log\sigma^2.\end{aligned}$$

Now subtract $\log \mathrm{E}\,X^2 = \log\sigma^2$ from both sides and observe that $h_b(\mu(t)) \ge 0$. Exponentiating both sides yields $M_G \le c(t)$. ∎

An immediate consequence of this corollary is that $(\alpha M_G)^{-1}$ is an upper bound to the MCQ coding gain $\Gamma_{\rm MCQ}$ (42). On one hand, a sparse source ($M_G \ll 1$) is a necessary condition for large MCQ coding gain, that is for the existence of a t^* such that $c(t^*) \ll 1$. On the other hand, if M_G is close to one, $\Gamma_{\rm MCQ}$ is necessarily small. The quantity $\Gamma_s = M_G^{-1}$ might be called *sample coding gain*, since it is the limit in sample size of the geometric mean of a sample divided by its arithmetic mean (see also Definition 3).

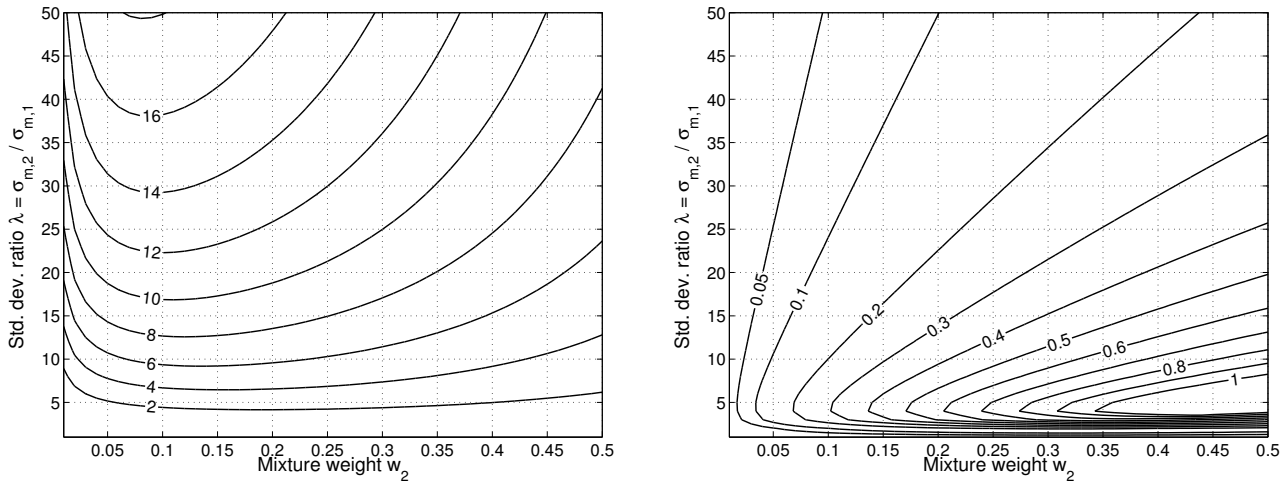
D. Examples

1) Coding Gain Loss for Gaussian Mixtures: If a transform outputs zero-mean Gaussian coefficients, such that each coefficient has one of just two distinct variances, the resulting marginal density will be a two-component Gaussian mixture (33). The largest coding gain would be achieved if both encoder and decoder knew the mixing variable S without needing extra rate. Then two codebooks matched to the variances could be used, like in a classical KLT water-filling solution. That situation corresponds exactly to the oracle lower bound (34) in Section VII-A, and the coding gain is simply the ratio from the Gaussian upper bound for the average variance to (34). If instead we mix the sources and apply MCQ, the resulting coding gain loss Δ_{CG} (at high rate) will be the ratio of the classical coding gain (41) to the MCQ coding gain (42), which is equal to the ratio of the high-rate upper bound (12) to the lower bound (34),

$$\Delta_{CG} = \frac{\Gamma_{\rm TC}}{\Gamma_{\rm MCQ}} = \frac{e^{2h_b(\mu^*)}\,[\sigma_0^*]^{2(1-\mu^*)}\,[\sigma_1^*]^{2\mu^*}}{\sigma_{m,1}^{2(1-w_2)}\sigma_{m,2}^{2w_2}}. \qquad (44)$$

Note that here the $[\sigma_i^*]^2$ are the variances of the sample classes in optimized MCQ (42), while the $\sigma^2_{m,s}$ are mixture component variances of the model (33). Fig. 9 shows contour plots of (a) the coding gain Γ_{TC} and (b) the coding gain loss Δ_{CG} (both in dB) for different ratios $\lambda^2 = \sigma^2_{m,2}/\sigma^2_{m,1}$ of the mixture variances and weights $w_2 = 1-w_1$ ($\lambda = 1$ yields the Gaussian pdf). Large λ and small w_2 lead to peaked densities; for example the wavelet coefficient mixture from Section VII has $\lambda \approx 30.9$ and $w_2 \approx 0.09$. From the graph, we see that these values correspond to a loss of about 2.5 dB, which can be verified by checking the distance between the high-rate bounds in Fig. 7.

The above definition of coding gain loss is based on the assumption that two distinct ($\lambda > 1$) Gaussian sources are mixed together even though they could be distinguished. That is, we are comparing a system where encoder and decoder know the variance (and the pdf) of each sample through some additional means costing no rate (like in traditional KLT transform coding) with a system where they know only the mixture pdf. However, in the case of a true mixture source neither the encoder nor the decoder know the variance of each sample, i.e. they do not known from which component source the sample originated. In that case the lower bound (34) is not tight for $\lambda > 1$, since the mixing random variable S is unknown, see also (35). Then a better definition of high-rate coding gain loss is the ratio of the high-rate upper bound to the Shannon lower bound (which asymptotically equals the



(a) (b)

Fig. 10. Magnitude classifying quantization (MCQ) of two-component Gaussian mixtures (GM) with component variances $\sigma^2_{m,1}$, $\sigma^2_{m,2}$, and weights $1-w_2$, w_2, respectively. (a) Contours of constant coding gain Γ_{SLB} (43) (in dB) for mixture source (equivalent to Shannon lower bound). (b) Coding gain loss $\Delta_{\mathrm{CG(SLB)}}$ (45) (in dB) relative to Γ_{SLB} for MCQ of the mixture.

distortion rate function):

$$\Delta_{\mathrm{CG(SLB)}} = \frac{\Gamma_{\mathrm{SLB}}}{\Gamma_{\mathrm{MCQ}}} = \frac{e^{2h_b(\mu^*)}\,[\sigma_0^*]^{2(1-\mu^*)}\,[\sigma_1^*]^{2\mu^*}}{\exp\left(2h(X) - \log(2\pi e)\right)}. \quad (45)$$

The GM differential entropy $h(X)$ has to be computed with numerical integration methods. Fig. 10 plots the coding gain Γ_{SLB} and the coding gain loss $\Delta_{\mathrm{CG(SLB)}}$ for that case (both in dB). The loss is remarkably low over a wide range of parameter values, which shows that the magnitude classification quantization approach is very effective for such sources. In this example, the optimal MCQ threshold t^* was always larger than the threshold for the maximum likelihood classification, $t_{\mathrm{ML}} = \sqrt{\log \lambda^2/(1-\lambda^{-2})}\sigma_{m,1}$. This is quite expected, since the goal of the classification is a tight distortion bound, not the optimal distinction of the two component sources.

2) Mixture versus Vector Coding Gain: The above example considered two-component Gaussian mixtures as models for sparse sources and compared different measures of coding gain. Here we will extend that model to N Gaussian components and exploit the simple relationship between their variances and the geometric mean of their mixture. This can be used to bound the coding gain of a Gaussian mixture as a function of the coding gain for the unmixed sources.

The goal is to compare the classical vector coding gain for N independent Gaussian sources, on the one hand, with the coding gain for a mixture source that outputs one of these N sources uniformly at random, on the other hand. For example, consider a transform that outputs N independent zero-mean Gaussian components. If we know the variance of each component, like in the KLT case, we can achieve the vector (transform) coding gain. If however only the distribution of the variances is known, then we can design a codebook for the corresponding scalar mixture source and still achieve the mixture coding gain. This is akin to a KLT-like transform for which the eigenvalues of the covariance matrix are known, but not their ordering. Intuitively, wavelet transforms lie between these two extremes, since e.g. coefficient variances are correlated across scales (but this also violates the independence assumption in the definition of coding gain).

Two results from Section VII-B will be useful. The logarithm of the geometric mean (lgm) of N variances,

$$\log G_N(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2) = \frac{1}{N}\sum_{i=1}^{N} \log \sigma_i^2, \quad (46)$$

differs only by a constant from the lgm of a Gaussian mixture (36) with component variances σ_i^2 and uniform weights $w_i = 1/N$. (Uniform weights are assumed for simplicity, but the following results can be extended to non-uniform weights.) Letting $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2$, the *vector* coding gain of N-dimensional Gaussian transform coding (41) relates to the lgm (46) as

$$\frac{1}{N}\sum_{i=1}^{N} \log \sigma_i^2 = \log(\sigma^2/\Gamma_{\mathrm{TC}}). \quad (47)$$

Now (46) can also be used to lower bound the mixture entropy $h(X)$ through (36) and (37), which leads to an upper bound on the *mixture* coding gain (43). Combining this with (47) yields $\Gamma_{\mathrm{SLB}} \leq \Gamma_{\mathrm{TC}}$, which simply means that mixing does not necessarily inflict a performance penalty. More interestingly, the same approach can be used with the upper bound on $h(X)$ in Corollary 10, which then yields a lower bound on Γ_{SLB} as a function of Γ_{TC}. If N is known, the second bound in (38) leads to a tighter lower bound, but only for large Γ_{TC} (the gap to the upper bound will be $20\log_{10} N$ [dB]).

Fig. 11 displays the upper and lower bounds for Gaussian mixture vs. vector coding gain. The lower curve thus limits the maximum performance loss (in dB) of a transform coding system that knows only the expected number of transform coefficients with a certain variance, but not their positions, compared to a system in which those positions are known (*a priori*, in the case of the KLT). The upper bound implies that the minimal performance loss is 0 dB.

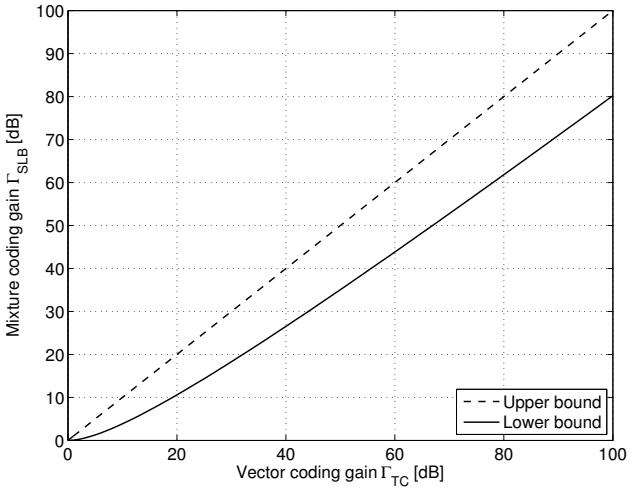


Fig. 11. Bounds for Gaussian mixture vs. vector coding gain.

VIII. Applications

The aim of this section is to give a brief overview of the rate distortion behavior of sparse sources in distributed coding (Wyner-Ziv) settings, as well as the relationship with compressed sensing.

A. Sparse Sources in Wyner-Ziv Settings

The Wyner-Ziv (WZ) problem [40], [41], [10, Sec. 14.9] considers pairs (X, Y) of dependent random variables and asks for the minimal rate $R^{\text{WZ}}_{X|Y}(D)$ required to describe the *source* X within average distortion D when *side information* Y is available only at the decoder. We limit our discussion to absolutely continuous X, Y and quadratic distortion measure $d(\hat{x}, x) = (\hat{x} - x)^2$. The WZ rate distortion function is given in [40], [41] as

$$R^{\text{WZ}}_{X|Y}(D) = \inf I(X; U|Y), \tag{48}$$

where the infimum is taken over all random variables U such that $Y \to X \to U$ form a Markov chain and there exists a function $f(U, Y) = \hat{X}$ such that $\mathrm{E}\, d(f(U, Y), X) \leq D$. It is lower-bounded by the corresponding conditional rate distortion function $R_{X|Y}(D)$, which is in turn lower-bounded by the conditional Shannon lower bound [34],

$$\begin{aligned} R^{\text{WZ}}_{X|Y}(D) \geq R_{X|Y}(D) &= \inf_{\{U \in \mathbb{R}:\ \mathrm{E}(U-X)^2 \leq D\}} I(X; U|Y) \\ &\geq h(X|Y) - \frac{1}{2} \log 2\pi e D, \end{aligned} \tag{49}$$

where U is a real-valued random variable. These lower bounds are asymptotically tight for the quadratic distortion measure [42]; in particular, the rate loss $R^{\text{WZ}}_{X|Y}(D) - R_{X|Y}(D)$ is zero for all D when (X, Y) are jointly Gaussian [41] and, more generally, when $X = Y + Z$ with Y independent from Z, where only Z needs to be Gaussian [43].

We further specialize to models where X and Y have zero mean and their difference can be modeled by an independent memoryless source, yielding the following two correlation models that are common in the WZ literature.

1) $X = Y + Z$, where Y, Z are Independent Memoryless Sources: The case of Y sparse and $Z \sim \mathcal{N}(0, \sigma_Z^2)$ is trivial, since [43] implies $R^{\text{WZ}}_{X|Y}(D) = R_{X|Y}(D) = \frac{1}{2} \log^+ \left(\frac{\sigma_Z^2}{D} \right)$, where $\log^+ x = \max\{\log x, 0\}$. More interesting is the case of sparse Z, since by [42] one has

$$R^{\text{WZ}}_{X|Y}(D) \doteq h(X|Y) - \frac{1}{2} \log 2\pi e D = h(Z) - \frac{1}{2} \log 2\pi e D, \tag{50}$$

where $\doteq$ denotes asymptotic equality for $D \to 0$ and the right-hand side is the SLB for $R_Z(D)$, the MSE rate distortion function for Z (this can also be shown directly using a "Gaussian forward test channel," see Fig. 4 in [41]). Thus all the presented techniques for bounding the asymptotic behavior of $R_Z(D)$ apply in this WZ case as well.

2) $Y = X + Z$, where X, Z are Independent Memoryless Sources: A Gaussian upper bound is obtained by bounding the rate $R_g(D) = I(X; U|Y)$ achieved with a Gaussian forward test channel, which consists in letting $U = \alpha(X + \Psi)$ with independent Gaussian noise Ψ, using an LMMSE reconstruction function $f(U, Y)$ and choosing $\mathrm{Var}(\Psi)$ and α such that the distortion constraint is met (see [41]). The bound is

$$R^{\text{WZ}}_{X|Y}(D) \leq R_g(D) \leq \frac{1}{2} \log^+ \left(\frac{\sigma^2_{X|Y}}{D} \right), \tag{51}$$

where the conditional variance $\sigma^2_{X|Y} = \frac{\sigma_X^2 \sigma_Z^2}{\sigma_X^2 + \sigma_Z^2}$, with equality on both sides if and only if X, Z are jointly Gaussian.

By inserting $h(X|Y) = h(X) - h(X + Z) + h(Z)$ in (49) and observing that $h(X + Z) \geq \max\{h(X), h(Z)\}$, one sees that $R^{\text{WZ}}_{X|Y}(D)$ is asymptotically upper-bounded by $\min\{R_X(D), R_Z(D)\}$. If $|h(X) - h(Z)| < \frac{1}{2} \log 2$, lower bounding $h(X + Z)$ with the entropy power inequality [10, Sec. 16.7] yields the slightly tighter asymptotical upper bound $\max\{R_X(D), R_Z(D)\} - \frac{1}{2} \log 2$. Upper bounds from the previous sections may be applied; however, if tighter bounds are desired, all three entropies $h(X), h(Z), h(X + Z)$ need to be bounded individually, which may require sharper tools than those presented.

Better bounds exist if Z is a Gaussian mixture of the form (33), with mixture components $Z_s \sim \mathcal{N}(0, \sigma_s^2)$ and weights w_s. By assuming that both the encoder and the decoder have access to the hidden mixing random variable S, one obtains the lower bound

$$\inf I(X; U|Y) \geq \inf \sum_s w_s I(X; U|X + Z_s),$$

where U is a random variable satisfying $\mathrm{E}(U - X)^2 \leq D$ as in (49). This infimum can be evaluated using rate allocation like in Section VII-A (a version of this bound for binary S first appeared in [44]). Asymptotically for $D \to 0$, this simplifies to

$$\begin{aligned} R^{\text{WZ}}_{X|Y}(D) &\dot{\geq} \sum_s w_s \left[h(X|X + Z_s) - h(X - U) \right] \\ &\geq \sum_s w_s h(X|X + Z_s) - \frac{1}{2} \log 2\pi e D, \end{aligned}$$

where $h(X|X+Z_s) = h(X) + h(Z_s) - h(X+Z_s)$ and $\dot{\geq}$ denotes asymptotic inequality for $D \to 0$. For Gaussian $X \sim \mathcal{N}(0, \sigma_X^2)$, this further reduces to

$$R_{X|Y}^{\mathrm{WZ}}(D) \dot{\geq} \frac{1}{2} \sum_s w_s \log \frac{\sigma^2_{X|X+Z_s}}{D}, \tag{52}$$

where the $\sigma^2_{X|X+Z_s} = \frac{\sigma_X^2 \sigma_s^2}{\sigma_X^2+\sigma_s^2}$ are the conditional variances given a single mixture component Z_s. A slightly tighter bound may be obtained by assuming that only the decoder has access to S, using techniques from [45], which are however unlikely to yield analytic expressions even in the Gaussian case.

An asymptotic upper bound for $D \to 0$ is obtained from $I(X;S|Y) = h(X|Y) - h(X|Y,S) = H(S|Y) - H(S|Y,X) \leq H(S)$ (F. Bassi, personal communication) as

$$\begin{aligned} R_{X|Y}^{\mathrm{WZ}}(D) &\doteq h(X|Y) - \frac{1}{2}\log 2\pi e D \\ &\leq \sum_s w_s h(X|X+Z_s) - \frac{1}{2}\log 2\pi e D + H(S), \end{aligned}$$

which for Gaussian $X \sim \mathcal{N}(0, \sigma_X^2)$ becomes

$$R_{X|Y}^{\mathrm{WZ}}(D) \dot{\leq} \frac{1}{2} \sum_s w_s \log \frac{\sigma^2_{X|X+Z_s}}{D} + H(S). \tag{53}$$

Whether this is sharper than the Gaussian upper bound (51) needs to be checked on a case-by-case basis. If X, Z are jointly Gaussian (and thus S constant), the asymptotic bounds (52) and (53) coincide and are equal to $R_{X|Y}^{\mathrm{WZ}}(D)$ *for all* $D \leq D_{\max} = \sigma^2_{X|X+Z}$. Clearly, (52) and (53) mirror (37) and (38) in the standard $R(D)$ case. The same comments on tightness made after (35) in Section VII-A apply by substituting (U, Y) for $\hat{X}$.

B. Connections with Compressed Sensing

We briefly outline how lossy coding of sparse sources is related to compressed sensing (CS) [3], [4] (see also the special issue [46]). A typical example of a CS problem is the compressive representation of a signal vector $\boldsymbol{x} \in \mathbb{R}^N$ of the form $\boldsymbol{x} = \Psi \boldsymbol{s}$, where Ψ is an orthonormal N-by-N matrix and $\boldsymbol{s} \in \mathbb{R}^N$ has at most K non-zero components (we say that $\boldsymbol{x}$ is strictly K-sparse with respect to Ψ). The problem is then to determine a sampling/compression mechanism for $\boldsymbol{x}$ without using the *sparsifying basis* Ψ at the encoder (e.g. for complexity reasons). CS typically involves sampling $\boldsymbol{x}$ using an M-by-N random *measurement matrix* Φ that is "fat" (i.e. has $M \ll N$) and has low coherence with Ψ. The key question concerns the number M of real-valued samples (the height of Φ) needed for the exact (lossless) reconstruction of all K-sparse signal vectors $\boldsymbol{s}$ with high probability. Compression is thus achieved in the sense of needing a number of samples M that may be much smaller than N. A *distributed* CS problem might consider a signal which is known to have a sparse difference with respect to a reference signal $\boldsymbol{y}$ (side information) available only at the decoder. This can be extended to multiple correlated signals, which may be composed of sparse and non-sparse components, and have to be sampled and encoded independently [47]. A related model, with correlated signals obtained by sparse filtering, is studied in [48].

In practice, the samples $\Phi \boldsymbol{x}$ must be quantized, say with R_{CS} bits each, if they are to be sent to a remote decoder. This implies some loss in the reconstruction of $\boldsymbol{s}$ (if it succeeds at all), which will depend on the total rate MR_{CS}. For benchmarking purposes, it is thus interesting to study an information-theoretic view of this *noisy* CS problem,[5] by considering the rate needed for approximate (lossy) reconstruction of almost all $\boldsymbol{s}$, i.e. the rate distortion behavior for an appropriate random model of $\boldsymbol{s}$. To simplify the analysis, one may consider asymptotically long sequences from a sparse memoryless source under MSE distortion measure, e.g., a Bernoulli-Gaussian spike with $p = \frac{K}{N}$ for modeling strictly sparse signals, or a Gaussian mixture model for sparse spikes with background noise. Notice that the CS and information-theoretic models differ in which quantities are considered random, which deterministic, and what reconstruction guarantees are given. The viewpoint here is that for a practical lossy coding system with an average distortion constraint, if the source can be modeled as random, the rate-distortion bounds will apply regardless whether CS is employed in the system or not. The work [12] is among the first to give sharper bounds on the $D(R)$ behavior of quantized CS of strictly sparse sources, but it does not provide a purely information-theoretic analysis framework.

When Ψ is assumed known at the encoder, the upper and lower bounds in this paper may be applied to appropriate random source models in order to benchmark the operational rate distortion behavior of practical quantized CS systems. This also extends to distributed scenarios like the model JSM-3 in [47], which can be related to the Wyner-Ziv setting $X = Y + Z$ mentioned above. The key is that knowing Ψ, the encoder can always obtain the signal $\boldsymbol{s}$ that is sparse in the standard basis, which can thus be modeled as a sparse source as outlined. The theoretical performance limits hold regardless whether a practical encoder uses Ψ or not. If $R(D)$ is the rate distortion function of the sparse source model, given target distortion D, any quantized CS system must satisfy $MR_{\mathrm{CS}} \geq NR(D)$ asymptotically for $N, M \to \infty$. This yields a simple trade-off between the CS sampling ratio M/N and the rate R_{CS} at which samples are quantized.

When Ψ is unknown at the encoder, two approaches may be thought of. One is to postulate the existence of an algorithm that finds a sparsifying basis Ψ knowing only the sparsity $\frac{K}{N}$ and the noisy measurements $\boldsymbol{y} = \Phi \boldsymbol{x} + \boldsymbol{z}$ (for work in this direction, see [49]). Then one may again assume that Ψ is known. The other approach is to consider Ψ as a side information random variable available only at the decoder and study this particular kind of Wyner-Ziv problem, as has recently been suggested in [50].

CS with quantized incoherent measurements may be viewed as a doubly non-adaptive coding scheme that is oblivious of both the sparsifying basis Ψ and the location of nonzero samples. When $\Psi = I_N$, this becomes a singly non-adaptive

[5]The quantization noise may be modeled by parallel noisy channels, $\boldsymbol{Y} = \Phi \boldsymbol{X} + \boldsymbol{Z}$, whose total capacity, assuming e.g. independent zero-mean AWGN channels, depends on the signal-to-noise ratios $\mathrm{E}(\Phi \boldsymbol{X})_i^2 / \mathrm{E}\, Z_i^2$. The total channel capacity upper bounds the total rate MR_{CS}.

scheme that may be implemented with a lossy block code. In [51], such a code has been constructed by combining a q-ary nested uniform scalar quantizer with a q-ary syndrome source code. The scheme works in a Wyner-Ziv setting if the nonzero values are bounded, while in the standard case without side information, one may introduce a compander to gain a little extra performance, which for a Bernoulli-Gaussian spike (Sec. III-B) asymptotically becomes

$$D(R_v) \simeq \frac{p}{12} 6\pi\sqrt{3}\sigma_v^2 2^{-2R_v},$$

where p is the probability of a spike, σ_v its variance and $R_v = \log_2 q$ the quantizer rate. The total rate is $R(R_v) \simeq h_b(p) + pR_v$, which is the same rate that a "nonlinear" *adaptive* code would need, see (6).

IX. Conclusions

Sparsity is the key to nonlinear approximation and compressed sensing. Work in these areas is generally more concerned with the number of real-valued samples required for achieving a certain approximation error or exact reconstruction, rather than with the rate distortion trade-off that is implicit when samples are quantized. This paper studied the rate distortion behavior of sparse memoryless sources modeling that situation. We proposed incomplete moments as a compressibility measure and used them to bound low- and high-rate $D(R)$. Furthermore, we introduced the geometric mean as a single-parameter compressibility measure and used it to bound asymptotic $R(D)$ via the entropy, and to compare different types of transform coding gain. Thus non-strict sparsity and lossy compression can be related in quantitative fashion. These results apply to the MSE distortion criterion, while for Hamming distortion we showed that $R(D)$ can be computed exactly in some cases and that it becomes almost linear for very sparse sources.

Appendix

A. Rate Distortion Function of a Discrete Memoryless Source (DMS)

Definition 8 (Rate distortion function of a DMS) *Let $X \sim P$ be a discrete random variable with alphabet $\mathcal{X}$, $Q_{\hat{X}|X}(k|j)$ a conditional distribution defining the discrete reconstruction random variable $\hat{X}$ with alphabet $\hat{\mathcal{X}}$, $P_{X,\hat{X}}(j,k) = P(j)Q(k|j)$ the corresponding joint distribution, and $\rho(x,\hat{x})$ a bounded non-negative single-letter distortion measure. The* average distortion *associated with $Q(k|j)$ is*

$$d(Q) = \sum_{j,k} P(j)Q(k|j)\rho(j,k). \tag{54}$$

*A conditional probability assignment Q satisfying $d(Q) \le D$ is called D-*admissible *and the set of all such Q is $Q_D = \{Q(k|j) : d(Q) \le D\}$. The average mutual information ("description rate") induced by Q is*

$$I(Q) = \sum_{j,k} P(j)Q(k|j) \log \frac{Q(k|j)}{Q(k)}, \tag{55}$$

where $Q(k) = \sum_j P(j)Q(k|j)$. The rate distortion function is defined as $R(D) = \min_{Q\in Q_D} I(Q)$.

This convex optimization problem can be solved with the method of Lagrange multipliers [16], [10, Sec. 13.7]. We start with the functional

$$J(Q) = I(Q) + \lambda d(Q) + \sum_j \nu_j \sum_k Q(k|j), \quad \lambda \ge 0,$$

where the last term comes from the constraint that $Q(k|j)$ is a proper conditional distribution, i.e. satisfies $\sum_k Q(k|j) = 1$. The minimizing conditional distribution is given by

$$Q(k|j) = \frac{Q(k)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}}. \tag{56}$$

The marginal $Q(k)$ has to satisfy the following $\hat{N} = |\hat{\mathcal{X}}|$ conditions:

$$\sum_j \frac{P(j)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}} = 1, \qquad \text{if } Q(k) > 0, \tag{57}$$

$$\sum_j \frac{P(j)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}} \le 1, \qquad \text{if } Q(k) = 0. \tag{58}$$

For a tentative solution, given by a marginal $Q(k)$ satisfying (57), it can be shown that the conditions (58) are necessary and sufficient to yield a point on the $R(D)$ curve, either directly as in [16, Theorem 2.5.2], or via the Kuhn-Tucker conditions [10, Sec. 13.7]. The solution is further simplified through the following theorem by Berger:

Theorem 14 *[16, Theorem 2.6.1] No more than $N = |\mathcal{X}|$ reproducing letters need be used to obtain any point on the $R(D)$ curve that does not lie on a straight-line segment. At most, $\hat{N} = N + 1$ reproducing letters are needed for a point that lies on a straight-line segment.*

B. Rate Distortion of Binary $(1, N)$ Sources

Proof of Proposition 1: The following relies heavily on the results summarized in Appendix A, where it is shown that $R(D)$ can be computed by solving a set of equations involving the marginal distribution $Q(k)$ on the reconstruction alphabet. The symmetry of the source distribution, $P(j) = 1/N$, $j = 1, \ldots, N$, suggests the following marginal distribution (with a slight abuse of notation):

$$Q = (q_0, q_1 = q_2 = \ldots = q_N = \frac{1-q_0}{N}). \tag{59}$$

Recall that the symbols $0, 1, \ldots, N$ correspond to $\mathbf{0}, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_N$, respectively. Let us first assume that $q_k > 0$ holds for all k. Then the $N + 1$ conditions (57) have to be met. We make the substitution $\beta = e^{-\lambda}$ and insert our $Q(k)$ into the equation, first for $k \ne 0$:

$$\frac{\beta^0}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} + \frac{(N-1)\beta^2}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} = \frac{1}{P(j)} = N,$$

which after some algebra becomes

$$q_0((N-1)\beta^2 - N\beta + 1) = 0. \tag{60}$$

For $k = 0$ we get almost the same equation:

$$\frac{N\beta^1}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} = \frac{1}{P(j)} = N,$$

which becomes

$$(1-q_0)((N-1)\beta^2 - N\beta + 1) = 0. \tag{61}$$

The solution $\beta = 1$ corresponds to the point $(0, D_{\max} = 1)$ in the (R, D) plane, which is achieved by setting $q_0 = 1$. Therefore the interesting solution is $\beta = 1/(N-1)$, which inserted into (56) yields

$$Q(k|j) = q_k(N-1)^{1-\rho(j,k)}. \tag{62}$$

Inserting (62) into (54) we get the average distortion $d(Q) = 1 - \frac{N-2}{N}(1-q_0)$ and from (55) the rate $I(Q) = \frac{N-2}{N}(1-q_0)\log(N-1)$. Noting that these hold for $q_0 > 0$, we combine them to eliminate q_0 and get

$$D(R) = 1 - \frac{R}{\log(N-1)} \qquad \text{for } R < \frac{N-2}{N}\log(N-1). \tag{63}$$

This proves the first part of (4). When R reaches its upper bound in (63), D reaches $2/N$ and we have $q_0 = 0$. At that point, (60) will be satisfied for all β. According to condition (58), (61) now becomes an inequality:

$$(N-1)\beta^2 - N\beta + 1 \geq 0. \tag{64}$$

This is satisfied by $\beta \geq 1$ or $\beta \leq \frac{1}{N-1}$, which is equivalent to $\lambda \geq \log(N-1)$. The first solution ($\beta \geq 1$) can be discarded, since $\beta = 1$ has already been handled and $\beta > 1$ implies $\lambda < 0$, which is ruled out. The conditional distribution parameterized by β is

$$Q(k|j) = \begin{cases} 0, & k = 0 \\ \frac{\beta^{\rho(j,k)}}{1+(N-1)\beta^2}, & k \neq 0 \end{cases} \tag{65}$$

As before, we put this into (54) to get $d(Q) = \frac{2(N-1)\beta^2}{1+(N-1)\beta^2}$ and into (55) yielding

$$I(Q) = \log N - \frac{(N-1)\beta^2}{1+(N-1)\beta^2}\log(N-1) - h_b\left(\frac{1}{1+(N-1)\beta^2}\right).$$

Eliminating β from the last two equations yields the second part of (4). ∎

C. Proof and Discussion of the Low-Rate Bound on $D(R)$

Proof of Theorem 5: The bound (15) itself is simply a Gaussian upper bound on the significant samples, where the rate is reduced by $h_b(\mu(t))$ to account for coding the positions of those samples, plus the variance $\sigma^2 - A(t)$ of the uncoded insignificant samples. It holds for all $t \geq 0$ and $R \geq 0$, but is non-trivial only for $R > h_b(\mu(t))$.

The main task is thus to derive the locally optimal rate (17). The rate trade-off is no longer between two codebooks, but between the side information $h_b(\mu(t))$ and the rate for significant coefficients R_1; hence it cannot be solved using water-filling. Our approach is to temporarily fix a threshold t and determine the rate $R^*(t)$ corresponding to the midpoint of the common tangent of two bounds $B_{lr}(t, R)$ and $B_{lr}(t+\Delta t, R)$, for $\Delta t \to 0$. The resulting point $(R^*(t), B_{lr}(t, R^*(t)))$ is a candidate member of the lower convex hull of the family of bounds (15) and thus locally optimal (around t). Optimality is only local, since another bound of the family (15) might lie strictly below this candidate point; see the remarks after the proof.

We take two curves of the family (15), say $B_{lr}(t, r)$ and $B_{lr}(t+\Delta t, r)$, and determine their common tangent by solving the following system of equations:

$$\tfrac{\partial}{\partial r}B_{lr}(t,r)\big|_{r=r_0} = \tfrac{\partial}{\partial r}B_{lr}(t+\Delta t,r)\big|_{r=r_1} = s \tag{66}$$

$$\frac{B_{lr}(t+\Delta t, r_1) - B_{lr}(t, r_0)}{r_1 - r_0} = s. \tag{67}$$

A necessary condition for this approach is that $B_{lr}(t, r)$ be continuously differentiable in r, such that tangents are well-defined. From $\frac{\partial}{\partial r}B_{lr}(t,r) = -2\frac{A(t)}{\mu(t)}\exp\left(-2\frac{r-h(\mu(t))}{\mu(t)}\right)$, we see that this is the case for t such that $\mu(t) > 0$. The conditions under which the system has no solutions are discussed in the remarks after the proof.

Using the partial derivative $\frac{\partial}{\partial r}B_{lr}$, we solve (66) for r_1:

$$r_1 = \tfrac{\mu(t+\Delta t)}{\mu(t)}[r_0 - h(\mu(t))] + h(\mu(t+\Delta t)) - \tfrac{1}{2}\mu(t+\Delta t)\ln\tfrac{A(t)\mu(t+\Delta t)}{A(t+\Delta t)\mu(t)}$$

and start inserting this solution into (67):

$$\begin{aligned} s &= \frac{B_{lr}(t+\Delta t, r_1) - B_{lr}(t, r_0)}{r_1 - r_0} \\ &= \frac{A(t+\Delta t)e^{-2\frac{r_0-h(\mu(t))}{\mu(t)}+\ln\frac{A(t)\mu(t+\Delta t)}{A(t+\Delta t)\mu(t)}} - A(t)e^{-2\frac{r_0-h(\mu(t))}{\mu(t)}}}{r_1 - r_0} \\ &\quad - \frac{[A(t+\Delta t) - A(t)]}{r_1 - r_0} \\ &= \frac{[\mu(t+\Delta t)-\mu(t)]\frac{A(t)}{\mu(t)}e^{-2\frac{r_0-h(\mu(t))}{\mu(t)}} - [A(t+\Delta t)-A(t)]}{r_1 - r_0} \\ &= \tfrac{\partial}{\partial r}B(t,r)\big|_{r=r_0} = -2\tfrac{A(t)}{\mu(t)}e^{-2\frac{r_0-h(\mu(t))}{\mu(t)}}, \end{aligned} \tag{68}$$

where the last equality is actually again (66). For our convenience we make the substitutions $\Delta\mu = \mu(t+\Delta t) - \mu(t)$, $\Delta A = A(t+\Delta t) - A(t)$ to obtain

$$\tfrac{A(t)}{\mu(t)}e^{-2\frac{r_0-h(\mu(t))}{\mu(t)}}[\Delta\mu + 2(r_1 - r_0)] - \Delta A = 0. \tag{69}$$

Now we let

$$\begin{aligned} y &= -2\tfrac{r_0}{\mu(t)}, \\ \alpha &= \tfrac{A(t)}{\mu(t)}e^{2\frac{h(\mu(t))}{\mu(t)}}, \\ \beta &= \Delta\mu + 2(r_1 - r_0) - 2\tfrac{r_0}{\mu(t)}\Delta\mu \\ &= \Delta\mu - 2\tfrac{\mu(t+\Delta t)}{\mu(t)}h(\mu(t)) + 2h(\mu(t+\Delta t)) \\ &\quad - \mu(t+\Delta t)\ln\tfrac{A(t)\mu(t+\Delta t)}{A(t+\Delta t)\mu(t)} \end{aligned} \tag{70}$$

so that (69) becomes

$$\alpha e^y(-\Delta\mu\, y + \beta) - \Delta A = 0. \tag{71}$$

At this point we assume $f(-t)+f(t)>0$ and hence by continuity $\Delta t>0$ implies $\Delta\mu<0$ and $\Delta A<0$.[6] Therefore we can divide (71) by $-\alpha\Delta\mu$ and get

$$e^y(y-\tfrac{\beta}{\Delta\mu})+\tfrac{\Delta A}{\alpha\Delta\mu}=0, \tag{72}$$

which can be solved using the Lambert W function:

$$y=\tfrac{\beta}{\Delta\mu}+W\left(-\tfrac{\Delta A}{\alpha\Delta\mu}e^{-\frac{\beta}{\Delta\mu}}\right). \tag{73}$$

Using the defining equation $W(x)e^{W(x)}=x$ it is easy to show that (73) actually solves (72). Before taking limits we resolve the question which is the correct branch of W to use. From (70) it is clear that we need negative real-valued solutions. The principal branch $W_0(x)$ has domain $[-1/e,\infty)$ and takes on values in $[-1,\infty)$, whereas the other real-valued branch $W_{-1}(x)$ has values in $(-\infty,-1]$. Since a more negative y will yield a tighter bound (see (70) and (15)), we pick the branch $W_{-1}(x)$. Its domain is $[-1/e,0)$, which implies that for a specific pdf $f(x)$ and threshold t, equation (73) might have no real solution, i.e. that no common tangent exists. That case will be analyzed in the remark following the proof.

Because $W_{-1}(x)$ is a continuous function, we may take the limit $\Delta t\to 0$ of the expressions appearing in its argument:

$$\begin{aligned}\lim_{\Delta t\to 0}\tfrac{\beta}{\Delta\mu}&=\lim_{\Delta t\to 0}1+2\tfrac{\mu(t)h(\mu(t+\Delta t))-\mu(t+\Delta t)h(\mu(t))}{\mu(t)[\mu(t+\Delta t)-\mu(t)]}\\&\quad+2\tfrac{\mu(t)h(\mu(t))-\mu(t)h(\mu(t))}{\mu(t)[\mu(t+\Delta t)-\mu(t)]}\\&\quad+\tfrac{\mu(t+\Delta t)\,\Delta t}{\mu(t+\Delta t)-\mu(t)}\left[\tfrac{\ln A(t+\Delta t)-\ln A(t)}{\Delta t}-\tfrac{\ln\mu(t+\Delta t)-\ln\mu(t)}{\Delta t}\right]\\&=1+2h'(\mu(t))-2\tfrac{h(\mu(t))}{\mu(t)}+\tfrac{\mu(t)A'(t)}{\mu'(t)A(t)}-1,\end{aligned} \tag{74}$$

$$\begin{aligned}\lim_{\Delta t\to 0}\tfrac{\Delta A}{\alpha\Delta\mu}&=\lim_{\Delta t\to 0}\tfrac{[A(t+\Delta t)-A(t)]\mu(t)}{[\mu(t+\Delta t)-\mu(t)]A(t)}e^{-2h(\mu(t))/\mu(t)}\\&=\tfrac{A'(t)\mu(t)}{A(t)\mu'(t)}e^{-2h(\mu(t))/\mu(t)}.\end{aligned} \tag{75}$$

By the definitions of μ and A we have $\mu'(t)=-f(-t)-f(t)$ and $A'(t)=(-f(-t)-f(t))t^2$, hence $\frac{\mu(t)A'(t)}{\mu'(t)A(t)}=\frac{\mu(t)}{A(t)}t^2=\gamma(t)$, as defined in (18). Inserting (74)–(75) into (73) gives

$$y=2h'(\mu(t))-2\tfrac{h(\mu(t))}{\mu(t)}+\gamma(t)+W\left(-\gamma(t)e^{-2h'(\mu(t))-\gamma(t)}\right) \tag{76}$$

and after inserting this into (70) and solving for r_0 we obtain (17). ∎

Remarks: As pointed out in the proof, for some values of the threshold t there might be no solution to the common tangent problem. This occurs when increasing t produces a bound $B_{lr}(t+\Delta t,r)$ that for all r is larger than $B_{lr}(t,r)$. Using a Taylor approximation

$$B_{lr}(t+\Delta t,r)\approx B_{lr}(t,r)+\tfrac{\partial}{\partial t}B(t,r)\Delta t$$

we see that there is no common tangent for those t satisfying

$$\min_r\tfrac{\partial}{\partial t}B(t,r)>0. \tag{77}$$

[6]If this assumption does not hold, we also have $\beta=0$ and thus (71) is always satisfied. This case corresponds to a pdf with symmetric holes in its support, that is, a mixture of two (or more) densities with non-overlapping supports. Picking the "critical" t actually separates the mixture components into two groups.

To find this minimum we differentiate $\frac{\partial}{\partial t}B$ with respect to r:

$$\begin{aligned}\tfrac{\partial^2}{\partial t\partial r}B(t,r)&=2e^{-2\frac{r-h(\mu(t))}{\mu(t)}}\\&\cdot\left[\frac{-\mu^2(t)A'(t)+2\mu'(t)A(t)\left(\frac{1}{2}\mu(t)-r+h(\mu(t))-\mu(t)h'(\mu(t))\right)}{\mu^3(t)}\right].\end{aligned}$$

After discarding the zero at $r=\infty$, the rate of the candidate minimum is found by setting the term in square brackets to 0:

$$r^*=\tfrac{1}{2}\mu(t)\left[1-\gamma(t)-2h'(\mu(t))+2\tfrac{h(\mu(t))}{\mu(t)}\right]. \tag{78}$$

To verify that it is indeed a (unique, thus global) minimum, we check the second derivative:

$$\tfrac{\partial^3}{\partial t\partial^2 r}B(t,r)\Big|_{r=r^*}=2e^{-2\frac{r^*-h(\mu(t))}{\mu(t)}}\left[4\tfrac{f(t)A(t)}{\mu^3(t)}\right]>0.$$

Now we insert (78) into (77) and obtain

$$\begin{aligned}\tfrac{\partial}{\partial t}B(t,r^*)&=\left[A'(t)+2\tfrac{A(t)\mu'(t)}{\mu(t)}\left(h'(\mu(t))+\tfrac{r^*-h(\mu(t))}{\mu(t)}\right)\right]\\&\quad\cdot e^{-2\frac{r^*-h(\mu(t))}{\mu(t)}}-A'(t)\\&=\tfrac{A(t)\mu'(t)}{\mu(t)}e^{\gamma(t)-1+2h'(\mu(t))}-A'(t)>0.\end{aligned} \tag{79}$$

If we divide (79) by $A'(t)<0$ (thus reversing the inequality) we see that (77) is equivalent to $e^{-1}e^{\gamma(t)+2h'(\mu(t))}<\gamma(t)$ and after another sign change we get

$$-\tfrac{1}{e}>-\gamma(t)e^{-\gamma(t)-2h'(\mu(t))}. \tag{80}$$

This is a reassuring result: condition (80), and thus (77), is true exactly if and only if the argument of the W function in (17) is less than $-1/e$, i.e. when there is no real-valued solution. Nevertheless, this condition alone does not guarantee that we find only convex hull points, since one of the constituent bounds $B_{lr}(t,r)$ might be *below* all others for a threshold t satisfying (77). In all examples we have studied, it happened to be the Gaussian bound $B_{lr}(0,r)$, if at all. For a Gaussian source, obviously no upper bound can be tighter than $B_{lr}(0,r)$, while for Laplacian sources all thresholds above a critical value yield slight improvements. For pdf's that are even more peaked around zero the critical threshold is (almost) zero. Informally, these "forbidden" threshold values mean that the reduction in the number of coded significant samples is not sufficient to offset the increased side information rate $h_b(\mu(t))$. From this reasoning it becomes evident that such t can only lie between 0 and $t_{0.5}$, with $\mu(t_{0.5})=0.5$. On the other hand, this means that the bound will be useful at low rates (larger t), which is exactly what is desired.

Corollary 15 *The low-rate bound* (16) *and the high-rate bound* (12) *coincide in proper non-boundary local extrema of* $c(t)$*, provided that* $f(-t)+f(t)>0$ *over* $supp(f)$*:*

$$\begin{aligned}R^*(t)=R_{\min}(t)&\iff B_{lr}(t,R^*(t))=B_{hr}(t,R_{\min}(t))\\&\iff c'(t)=0,\end{aligned} \tag{81}$$

where $c(t)$ *is defined in* (13) *and* $R_{\min}(t)$ *in* (11).

Proof: First we study the last equation. The derivative of $c(t)$ is

$$\begin{aligned} c'(t) &= c(t)\Big[\mu'(t)\ln\tfrac{\sigma_1^2(t)}{\sigma_0^2(t)} \\ &\qquad + A'(t)\left(\tfrac{1}{\sigma_1^2(t)}-\tfrac{1}{\sigma_0^2(t)}\right)+2\mu'(t)h'(\mu(t))\Big] \\ &= -(f(-t)+f(t))c(t)\Big[\ln\tfrac{\sigma_1^2(t)}{\sigma_0^2(t)} \\ &\qquad + t^2\left(\tfrac{1}{\sigma_1^2(t)}-\tfrac{1}{\sigma_0^2(t)}\right)+2h'(\mu(t))\Big]. \end{aligned} \tag{82}$$

Since we assumed $f(-t)+f(t)>0$ and have $c(t)>0$ by definition, the term in square brackets has to be zero for a proper local extremum. Since the domain $t\in[0,\infty)$ is half open, a possible boundary minimum at $t=0$ has to be inspected separately. (The same applies to the right boundary $t_{\max}$ if the support is bounded.) Now we inspect the middle equation:

$$\begin{aligned} B_{lr}(t,R^*(t)) &= A(t)\exp\Big[2h'(\mu(t))+\gamma(t)+ \\ &\qquad W_{-1}(-\gamma(t)e^{-2h'(\mu(t))-\gamma(t)})\Big]+\sigma^2-A(t) \\ &= B_{hr}(t,R_{\min}(t)) = \sigma_0^2(t) = \tfrac{\sigma^2-A(t)}{1-\mu(t)}, \end{aligned}$$

which after taking the logarithm is equivalent to

$$W_{-1}(-\gamma(t)e^{-2h'(\mu(t))-\gamma(t)}) = -2h'(\mu(t))-\gamma(t)-\ln\tfrac{\sigma_1^2(t)}{\sigma_0^2(t)}. \tag{83}$$

Inserting this into the defining equation $W(x)e^{W(x)}=x$ we get

$$\begin{aligned} \left(-2h'(\mu(t))-\gamma(t)-\ln\tfrac{\sigma_1^2(t)}{\sigma_0^2(t)}\right)\tfrac{\sigma_0^2(t)}{\sigma_1^2(t)}e^{-2h'(\mu(t))-\gamma(t)} = \\ -\gamma(t)e^{-2h'(\mu(t))-\gamma(t)}, \end{aligned} \tag{84}$$

which is equivalent to

$$\left[\ln\tfrac{\sigma_1^2(t)}{\sigma_0^2(t)}+2h'(\mu(t))+\gamma(t)\Big(1-\tfrac{\sigma_1^2(t)}{\sigma_0^2(t)}\Big)\right]e^{-2h'(\mu(t))-\gamma(t)}=0. \tag{85}$$

Observing that $\gamma(t)=\frac{t^2}{\sigma_1^2(t)}$ shows that the term in brackets is equal to the bracketed term in (82), so that (85) implies $c'(t)=0$. Finally, the first expression in (81) is

$$\begin{aligned} 0 &= R^*(t)-R_{\min}(t) \\ &= -\tfrac{\mu(t)}{2}\left[2h'(\mu(t))+\gamma(t)+W_{-1}\left(-\gamma(t)e^{-2h'(\mu(t))-\gamma(t)}\right)\right] \\ &\qquad -\tfrac{\mu(t)}{2}\ln\tfrac{\sigma_1^2(t)}{\sigma_0^2(t)}. \end{aligned} \tag{86}$$

Since t is not allowed to be on the right support boundary, we can divide (86) by $\mu(t)>0$ and after rearranging terms we get exactly equation (83). Thus also the first condition is equivalent to $c'(t)=0$. ∎

Remark: For a fixed threshold t, by definition of $R_{\min}$ the point $B_{hr}(t,R_{\min}(t))$ is the switch-point between the low-rate bound and the high-rate bound, that is for all $R>R_{\min}$ the high-rate bound is tighter. If now the two bounds are optimized ("best R for given t"), Corollary 15 comes as no big surprise. In the interesting cases, when $c(t)$ has a single local (thus global) minimum at $t_0>0$, the consequence is that for $t>t_0$ ($R<R_{\min}(t_0)$) the low-rate bound will be tighter, and for $R\geq R_{\min}(t_0)$ the high-rate bound will take over. In the less interesting cases such as the Gaussian, $c(t)$ is minimal at $t_0=0$ and takes on a global maximum for some $t_0>0$. At low rates the bound (16) is again tighter; it becomes looser up to $R_{\min}(t_0)$, while from that rate on (12) will be the loosest bound. So far we have found no examples of densities that lead to multiple local extrema of $c(t)$.

D. Maximum Entropy given Variance and Geometric Mean

Proof of Proposition 9:

The proof relies on the method for obtaining maximum entropy distributions outlined in [10, Ch. 11]. The goal is to maximize the entropy $h(f)$ over all probability densities f satisfying

$$\begin{array}{ll} 1) & f(x)\geq 0,\quad x\in\mathbb{R}, \\ 2) & \int f(x)\,\mathrm{d}x = 1, \\ 3) & \int f(x)\log|x|\,\mathrm{d}x = \theta, \\ 4) & \int f(x)x^2\,\mathrm{d}x = \sigma^2, \end{array} \tag{87}$$

where all integrals are over $\mathbb{R}$. Using calculus of variations, it can be shown that the maximizing density has the form

$$f(x) = e^{\lambda_0-1}|x|^{\lambda_1}e^{\lambda_2 x^2}, \tag{88}$$

where $\lambda_0,\lambda_1,\lambda_2$ are chosen such that f satisfies the constraints 2, 3, 4 in (87). Using an information inequality, it can then be shown that if there exists an f^* of the form (88) satisfying (87), then it is the unique maximizer over all densities satisfying (87) [10, Thm. 11.1.1]. Thus we need only prove that (29) satisfies the constraints (87).

The normalization constraint $\int f(x)\,\mathrm{d}x=1$ is satisfied if and only if $e^{\lambda_0-1}=(-\lambda_2)^{(\lambda_1+1)/2}/\Gamma((\lambda_1+1)/2)$. Furthermore, for the integral (and higher moments) to converge, we must have $\lambda_2<0$. Inserting the above expression for λ_0 into (88) yields the second moment $\int f(x)x^2\,\mathrm{d}x=-(\lambda_1+1)/(2\lambda_2)$, which satisfies the corresponding constraint if and only if $\lambda_2=-(\lambda_1+1)/(2\sigma^2)$. To simplify expressions, we substitute $\lambda_1=u-1$. The condition $\lambda_2<0$ thus becomes $u>0$.

We need to show that $\mathrm{E}\log|X|$ is monotone increasing in u, so that the mapping between θ and u defined by (30) is one-to-one. By Jensen's inequality we have $\mathrm{E}\log|X|\leq\frac{1}{2}\log\mathrm{E}\,X^2=\log\sigma$. Let $v=u/2$ and $\phi(v)=2(\mathrm{E}\log|X|-\frac{1}{2}\log\mathrm{E}\,X^2)=\psi(v)-\log v$. Using a standard integral representation for $\psi(v)$ [32, 8.361.3] we obtain

$$\phi(v) = -\frac{1}{2v}-2\int_0^\infty\frac{t\,\mathrm{d}t}{(t^2+v^2)(e^{2\pi t}-1)},\qquad v>0. \tag{89}$$

The first derivative,

$$\phi'(v) = \frac{1}{2v^2}+4\int_0^\infty\frac{vt\,\mathrm{d}t}{(t^2+v^2)^2(e^{2\pi t}-1)}, \tag{90}$$

is strictly positive for $v>0$, so $\mathrm{E}\log|X|$ is indeed monotone increasing. By bounding the integral in (89), one can further show that $\lim_{u\to\infty}\mathrm{E}\log|X|=\log\sigma$, which means that all admissible constraints $\theta\leq\log\sigma$ can be satisfied. Thus there is a unique $\lambda_1=u-1$ satisfying constraint 3 in (87), which together with σ^2 in turn determines λ_2 (satisfying constraint 4) and finally λ_0 (satisfying constraint 2). ∎

ACKNOWLEDGMENTS

The authors wish to thank Emre Telatar for helpful discussions, as well as the anonymous reviewers and the associate editor for their valuable suggestions and comments, which greatly helped improving the presentation of this material.

REFERENCES

[1] D. L. Donoho, M. Vetterli, R. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 2435–2476, Oct. 1998.

[2] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Proc.*, vol. 50, no. 6, pp. 1417–1428, Jun. 2002.

[3] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[4] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[5] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 83–91, Mar. 2008.

[6] M. Vetterli, "Wavelets, approximation, and compression," *IEEE Signal Processing Mag.*, vol. 18, pp. 59–73, Sep. 2001.

[7] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[8] A. Cohen and J.-P. D'Alès, "Nonlinear approximation of random functions," *SIAM J. Appl. Math.*, vol. 57, no. 2, pp. 518–540, Apr. 1997.

[9] S. Mallat and F. Falzon, "Analysis of low bit rate image transform coding," *IEEE Trans. Signal Proc.*, vol. 46, pp. 1027–1042, Apr. 1998.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.

[11] A. Cohen, I. Daubechies, O. Guleryuz, and M. Orchard, "On the importance of combining wavelet-based nonlinear approximation with coding strategies," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1895–1921, Jul. 2002.

[12] A. K. Fletcher, S. Rangan, and V. K Goyal, "On the rate-distortion performance of compressed sensing," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, Apr. 15–20 2007, pp. III–885–III–888.

[13] E. P. Simoncelli, "Statistical modeling of photographic images," in *Handbook of Video and Image Processing*, 2nd ed., A. C. Bovik, Ed. Academic Press, 2005, pp. 431–442. [Online]. Available: http://www.cns.nyu.edu/pub/eero/simoncelli05a-preprint.pdf

[14] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, Jul. and Oct. 1948, also in *Claude Elwood Shannon: Collected Papers*, N.J.A Sloane and A.D. Wyner, Eds., IEEE Press 1993, pp. 5–83.

[15] ——, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Conv. Rec.*, vol. 7, 1959, pp. 142–163, also in *Claude Elwood Shannon: Collected Papers*, N.J.A Sloane and A.D. Wyner, Eds., IEEE Press 1993, pp. 325–350.

[16] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.

[17] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 2026–2031, Nov. 1994.

[18] C. Weidmann, "Oligoquantization in low-rate lossy source coding," Ph.D. dissertation, EPFL, Lausanne, Switzerland, July 2000.

[19] M. Pinsker, *Information and Information Stability of Random Variables and Processes*. New York: Holden-Day, 1964.

[20] H. Rosenthal and J. Binia, "On the epsilon entropy of mixed random variables," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 1110–1114, Sep. 1988.

[21] A. György, T. Linder, and K. Zeger, "On the rate-distortion function of random vectors and stationary sources with mixed distributions," *IEEE Trans. Inform. Theory*, vol. IT-45, pp. 2110–2115, Sep. 1999.

[22] B. Bénichou and N. Saito, "Sparsity vs. statistical independence in adaptive signal representations: A case study of the spike process," in *Beyond Wavelets*, ser. Studies in Computational Mathematics, G. V. Welland, Ed. Academic Press, 2003, vol. 10, ch. 9, pp. 225–257. [Online]. Available: http://www.math.ucdavis.edu/~saito/publications/

[23] M. O. Lorenz, "Methods of measuring the concentration of wealth," *Publications of the American Statistical Association*, vol. 9, no. 70, pp. 209–219, Jun. 1905.

[24] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Inform. Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.

[25] G. Hardy, J. Littlewood, and G. Pólya, *Inequalities*, 2nd ed. Cambridge University Press, 1952.

[26] D. J. Sakrison, "Worst sources and robust codes for difference distortion measures," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 301–309, May 1975.

[27] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 676–683, Sep. 1968.

[28] A. M. Gerrish and P. M. Schultheiss, "Information rates of non-Gaussian processes," *IEEE Trans. Inform. Theory*, vol. IT-10, pp. 265–271, Oct. 1964.

[29] A. D. Wyner, "An upper bound on the entropy series," *Inform. Contr.*, vol. 20, pp. 176–181, 1972.

[30] E. H. Lieb and M. Loss, *Analysis*, 2nd ed. American Mathematical Society, 2001.

[31] J. Cheng, T.-K. Huang, and C. Weidmann, "New bounds on the expected length of optimal one-to-one codes," *IEEE Trans. Inform. Theory*, vol. 53, no. 5, pp. 1884–1895, May 2007.

[32] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 5th ed. New York: Academic Press, 1994.

[33] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Proc.*, vol. 46, pp. 886–902, Apr. 1998.

[34] R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 480–489, Jul. 1973.

[35] A. Hjørungnes and J. M. Lervik, "Jointly optimal classification and uniform threshold quantization in entropy constrained subband image coding," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1997, pp. 3109–3112.

[36] R. M. Gray, "Gauss mixture vector quantization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, May 2001, pp. 1769–1772.

[37] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. Royal Statistical Society, Series B*, vol. 36, no. 1, pp. 99–102, 1974.

[38] V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Processing Mag.*, vol. 18, no. 5, pp. 9–21, Sep. 2001.

[39] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

[40] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the receiver," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.

[41] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder-II: General sources," *Inform. Contr.*, vol. 38, pp. 60–80, 1978.

[42] R. Zamir, "The rate loss in the Wyner-Ziv problem," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2073–2084, Nov. 1996.

[43] S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extension to the side information case," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1181–1203, May 2003.

[44] F. Bassi, M. Kieffer, and C. Weidmann, "Source coding with intermittent and degraded side information at the decoder," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Mar. 30 – Apr. 4, 2008.

[45] C. T. K. Ng, C. Tian, A. J. Goldsmith, and S. Shamai (Shitz), "Minimum expected distortion in Gaussian source coding with uncertain side information," in *Proc. Information Theory Workshop (ITW)*, Lake Tahoe, CA, USA, Sep. 2–6 2007, pp. 454–459.

[46] R. G. Baraniuk, E. Candès, R. Nowak, and M. Vetterli (guest editors), "Sensing, sampling, and compression," *IEEE Signal Processing Mag.*, vol. 25, no. 2, Mar. 2008.

[47] M. Wakin, M. Duarte, S. Sarvotham, D. Baron, and R. Baraniuk, "Recovery of jointly sparse signals from few random projections," in *Proc. Neural Information Processing Systems (NIPS)*, Dec. 2005.

[48] A. Hormati, O. Roy, Y. M. Lu, and M. Vetterli, "Distributed sampling of signals linked by sparse filtering: Theory and applications," *IEEE Trans. Signal Proc.*, vol. 58, no. 3, pp. 1095–1109, 2010.

[49] S. Gleichman and Y. C. Eldar, "Blind compressed sensing," Feb. 2010, submitted to IEEE Trans. Inform. Theory. [Online]. Available: http://arxiv.org/abs/1002.2586

[50] V. K. Goyal, A. K. Fletcher, and S. Rangan, "Compressive sampling and lossy compression," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 48–56, Mar. 2008.

[51] C. Weidmann, F. Bassi, and M. Kieffer, "Practical distributed source coding with impulse-noise degraded side information at the decoder," in *Proc. EUSIPCO*, Lausanne, Switzerland, Aug. 2008.