OLIGOQUANTIZATION IN LOW-RATE LOSSY SOURCE CODING

THÈSE N° 2234 (2000)

PRÉSENTÉE AU DÉPARTEMENT DE SYSTÈMES DE COMMUNICATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Claudio WEIDMANN

Ingénieur électricien diplômé EPF
de nationalité suisse et originaire d'Adlikon (ZH)

acceptée sur proposition du jury:

Prof. M. Vetterli, directeur de thèse
Prof. J. C. Kieffer, rapporteur
Prof. H.-A. Loeliger, rapporteur
Prof. E. Telatar, rapporteur

Lausanne, EPFL
2002

Abstract

The theory of high rate lossy source coding is well developed both with respect to the practice of quantization and its fundamental rate distortion limits. But many modern compression systems for natural signals such as audio and images operate at lower rates, which are not covered by these theories. Often these systems are based on a signal transform followed by a nonlinear processing step that reduces the number of coefficients that are actually quantized. Both this nonlinearity and the non-Gaussianity of the involved signals rule out using the theory of (high rate) linear transform coding to analyze such compression algorithms. The lossy source coding theorem could be used, but it is not the right tool "to see what happens". Another way of getting a better understanding of these systems is to look at them as a form of nonlinear approximation of individual signals. While partly avoiding the need for statistical signal models that are hard to come by, this approach still relies on high rate results for the coefficient quantization.

In this thesis we take a more information theoretic approach, whose main result is a new low-rate upper bound on the distortion rate function. In order to obtain that result, the problem had to be simplified substantially. To begin with, only i.i.d. memoryless sources with a mean square error distortion measure are considered. Instead of bounding the distortion rate function $D(R)$ directly, we derive upper bounds on the optimum performance of a certain class of quantizers, which by definition upper bound also $D(R)$. Inspired by the observation that many of those compression systems quantize only very *few* coefficients at low rates, we called this the class of *oligo*quantizers. In particular, if the coefficients have a peaked unimodal probability density, a simple magnitude thresholding operation is very effective at deciding which coefficients should be quantized. It turns out that peaked unimodal densities are also the most interesting ones, not just because they appear in many transform coding systems, but also because they have a $D(R)$ behavior that is very different from a Gaussian. Thus the emphasis of this work will be on oligoquantization based on scalar thresholding. Actually we will show that the principle of oligoquantization does not generalize well to higher dimensions.

On a more applied note, we analyze several models for sparse transform coefficients with the aid of the low rate bound. For the special case of sparse coefficient vectors and Hamming distortion measure, analytic expressions for $R(D)$ are obtained. A high rate bound that complements its low rate counterpart is also derived. Finally,

we show how these bounds can be computed directly from a set of source samples, without first estimating the underlying probability density. These empirical bounds can be a useful complement to the rate distortion function computed with Blahut's algorithm, since the latter is hard to use at very low or high rates.

Zusammenfassung

Die Theorie der verlustbehafteten Quellcodierung mit hoher Rate ist gut entwickelt, sowohl in Bezug auf die praktischen Aspekte der Quantisierung, als auch betreffs der durch die Ratenverzerrungstheorie gesetzten grundlegenden Schranken. Aber viele moderne Kompressionssysteme für natürliche Signale wie Audio und Bilder arbeiten mit niedrigeren Raten, die nicht durch diese Theorien abgedeckt werden. Häufig basieren diese Systeme auf einem Signaltransformation gefolgt von einem nichtlinearen Verarbeitungsschritt, der die Anzahl Koeffizienten verringert, die effektiv quantisiert werden. Sowohl diese Nichtlinearität als auch der nicht-Gaussische Charakter der interessierenden Signale verunmöglichen es, solche Kompressionsalgorithmen mittels der Theorie der (hochratigen) Transformationskompression zu analysieren. Das Raten-Verzerrungs-Theorem könnte verwendet werden, aber es ist nicht das geeignete Hilfsmittel um zu verstehen warum solche Methoden derart gut funktionieren. Ein anderer Ansatz betrachtet diese Algorithmen als Methoden zur nichtlinearen Approximation von deterministischen Signalen. Obwohl dadurch auf schwer beizukommende statistische Signalmodelle teilweise verzichtet werden kann, verwendet dieser Ansatz immer noch die Resultate für hochratige Quantisierung.

In der vorliegenden Dissertation beschreiten wir einen mehr informationstheoretischen Zugang, dessen Hauptergebnis eine neue obere Schranke der Ratenverzerrungsfunktion für niedere Raten ist. Um dieses Resultat zu erreichen mussten wir das Problem stark vereinfachen. Einerseits beschränken wir uns auf gedächtnisfreie Quellen und quadratisches Verzerrungsmass. Andererseits arbeiten wir nicht direkt mit der Ratenverzerrungsfunktion $D(R)$, sondern mit oberen Schranken auf der operationellen Ratenverzerrungsfunktion einer bestimmten Klasse von Quantisierungsverfahren. Definitionsgemäss sind diese Schranken auch obere Schranken auf $D(R)$. Inspiriert durch die Beobachtung, dass viele dieser Kompressionsverfahren bei niedrigen Raten nur *wenige* Koeffizienten effektiv quantisieren, tauften wir diese Klasse von Verfahren "*Oligo*quantisierer". Insbesondere wenn die Koeffizienten eine zugespitzte unimodale Wahrscheinlichkeitsdichte haben, genügt eine einfache Schwellenoperation um zu bestimmen, welche Koeffizienten quantisiert werden sollen. Es stellt sich heraus, dass diese unimodalen Wahrscheinlichkeitsdichten am interessantesten sind, nicht nur weil sie in vielen Kompressionsverfahren auftreten, sondern auch weil sie ein $D(R)$-Verhalten haben, dass sich stark von dem Gaussischer Quellen unterscheidet. Daher liegt der Schwerpunkt dieser Arbeit in Oligoquantisierern, die auf skalarer Schwel-

lendiskriminierung beruhen. Tatsächlich werden wir zeigen, dass sich das Prinzip der Oligoquantisierung schlecht auf höhere Dimensionen verallgemeinern lässt.

Im Hinblick auf Anwendungen untersuchen wir verschiedene Modelle für Transformationskoeffizienten mit Hilfe der neuen oberen Schranke. Für einen Spezialfall mit Hamming-Verzerrung geben wir exakte parametrische Ausdrücke für $R(D)$ an. Die Schranke für niedere Raten ergänzen wir mit einer für hohe Raten. Schliesslich zeigen wir, wie diese Schranken direkt aus einer Stichprobe der Quelle berechnet berechnet werden können, ohne vorher die Wahrscheinlichkeitsdichte der Quelle schätzen zu müssen. Diese empirisch bestimmten Schranken sind eine willkommene Ergänzung zu den Ratenverzerrungsfunktionen, die mit dem Blahut-Algorithmus berechnet werden können. Letzterer kann nämlich bei sehr niedrigen oder hohen Raten nur beschränkt eingesetzt werden.

Acknowledgements

Long before I considered becoming myself one of those bizarre research assistants wearing Birkenstocks at work (I still don't), I had the luck to follow the lectures of Jim Massey at ETH Zurich. He is to be held responsible for my interest in information theory, which ultimately led to this thesis. Thank you, Jim, for your excellent teaching.

My biggest thanks go to my advisor Martin Vetterli, who gave me his support through all ups and downs of my thesis years. His openness to research off the beaten paths gave me almost unlimited freedom in my work. And his stimulating suggestions often helped me getting back on track, without spoiling the freedom and fun of research. I also wish to thank the committee members, John Kieffer, Andi Loeliger, Emre Telatar and Ruediger Urbanke, for reading and accepting my thesis. They had just a few weeks in early summer to read it, which makes their efforts twice as valuable. Furthermore, I gratefully acknowledge a fellowship of the ETH Council, which provided financial support for the largest part of my thesis years.

Doing research would have been impossible without the great atmosphere at the LCAV. David, Jérôme, Michael, Paolo, Pier Luigi, Pina and all other colleagues: thank you for the good times we had in and out of the lab. Special thanks go to my officemate Jérôme for filing my thesis while I was already at a conference in Sorrento. Thinking of conferences and workshops, which are an essential source of inspiration and motivation, I would like to thank the friends and colleagues who shared these experiences (and also hotel rooms and beers) with me, in particular Jossy, Vivek and Zsolt. I am especially grateful to Jossy for our collaboration on arithmetic channel coding, which was and is very enriching from many points of view. My thanks also go to the students whose projects I supervised.

Finally, I thank my parents and family, especially my "second family" in Küsnacht, for supporting me and putting up with my quirks and moods. Thank you, Didi, for everything and the rest.

The very last thanks go to *you* for reading till here. Though I don't offer any rewards for finding typos and mistakes, please consider reading on ...

Contents

Chapter 1

Introduction

The main motivation for writing the present thesis came from the astonishingly good low-rate performance of modern image compressors, wavelet-based and not. Their operational distortion rate curve has a steep decay at low rates, before it turns into the asymptotic -6 dB/bit slope for higher rates. The key observation to be made is that at low rates, only very few coefficients are actually quantized, and all the others are set to zero (which is also quantization, of course). This fact is illustrated in Figure 1.1, which shows both the distortion as a function of rate and the number of coefficients used to achieve that distortion. Prompted by how *few* coefficients are actually quantized, we coined the name *oligo*quantization for this situation.[1]

The goal we set out for this thesis is to analyze oligoquantization to obtain a better understanding of low-rate lossy compression. In the remainder of this introduction, we will briefly recall some of the basic concepts of quantization and rate distortion theory.

1.1 Quantization and Rate Distortion Theory

The following summary contains information that is found in the standard references on the subjects. For quantization, these are [2, 19, 20, 21, 29, 35] in no particular order, and for rate distortion theory we refer to [1, 10, 27, 44, 45]. The IEEE Transactions on Information Theory had a special issue on Quantization in May 1982; also the 50'th anniversary commemorative issue of October 1998 contains two related articles [3, 23].

A quantizer $Q(\boldsymbol{x})$ maps the whole space $\mathbb{R}^k$ to a countable set of *reproduction values* $\mathcal{Y} = \{\boldsymbol{y}_i : i \in \mathcal{I}\}$, where $\mathcal{I}$ is called the *index set* (often $\mathcal{I} = \mathbb{N}$ or $\mathbb{N}_0$). Therefore we can define a quantizer as a set of *cells* $\mathcal{S} = \{S_i \subseteq \mathbb{R}^k : i \in \mathcal{I}\}$ with their corresponding reproduction values, such that $Q(\boldsymbol{x}) = \boldsymbol{y}_i$ for $\boldsymbol{x} \in S_i$. Using an

[1] Compare with *oligo*mineral water which contains particularly *few* minerals ...

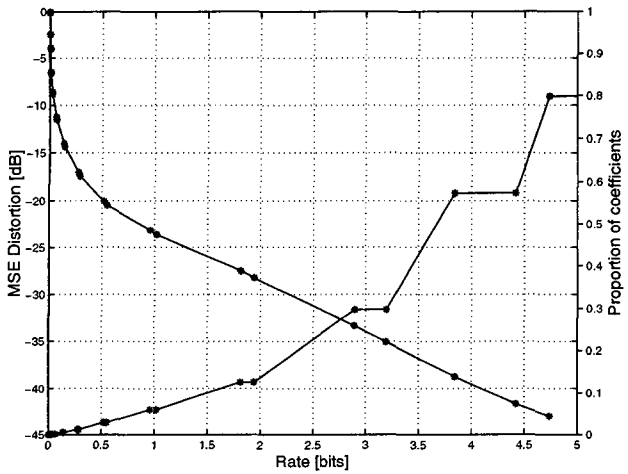

Figure 1.1: Typical wavelet transform image coder performance: operational distortion rate function (decreasing curve, left scale) and proportion of coefficients used (increasing curve, right scale).

indicator function we can also write

$$Q(\boldsymbol{x}) = \sum_i \boldsymbol{y}_i 1_{\boldsymbol{x} \in S_i}. \tag{1.1}$$

The quality of a quantizer is measured by defining a distortion measure $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$ between a sample $\boldsymbol{x}$ and its reproduction $\hat{\boldsymbol{x}}$. In this thesis we will only use the squared error:

$$d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2. \tag{1.2}$$

If the quantized samples are from a random source with probability density $f_{\boldsymbol{X}}(\boldsymbol{x})$, the average quantizer rate per dimension is

$$R(Q) = \frac{1}{k} H(Q(\boldsymbol{X})) = -\frac{1}{k} \sum_i p_i \log p_i \tag{1.3}$$

with $p_i = \Pr\{\boldsymbol{X} \in S_i\} = \int_{S_i} f_{\boldsymbol{x}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$. The average squared error distortion per dimension is

$$D(Q) = \frac{1}{k} \mathrm{E}[d(\boldsymbol{X}, Q(\boldsymbol{X}))] = \frac{1}{k} \sum_i p_i \int_{S_i} \|\boldsymbol{x} - Q(\boldsymbol{x})\|^2 f_{\boldsymbol{x}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{1.4}$$

For the mean squared error (MSE) distortion measure it can be shown that the lowest distortion is achieved if each cell is represented by its conditional mean or *centroid*:

$$\boldsymbol{y}_i = \mathrm{E}[\boldsymbol{X}|\boldsymbol{X} \in S_i] = \frac{1}{p_i} \int_{S_i} f_{\boldsymbol{x}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{1.5}$$

The tradeoff between rate and distortion may be quantified by the *k-dimensional operational distortion rate function* $\delta_k(R)$, which is the least distortion achievable by any k-dimensional quantizer of rate R or less:

$$\delta_k(R) = \inf_{Q:R(Q)\leq R} D(Q). \tag{1.6}$$

If we further optimize over the dimensionality, we get the *operational distortion rate function*:

$$\bar{\delta}(R) = \inf_k \delta_k(R). \tag{1.7}$$

This is the point where rate distortion theory enters the scene: while quantization uses a deterministic mapping $Q(x)$, rate distortion theory uses a stochastic mapping that for each input value x is defined by a conditional probability density $q_{Y|X}(y|x)$.[2] The distortion is defined as

$$D(q) = \mathrm{E}[d(X,Y)] \tag{1.8}$$

and the *information rate* is the mutual information

$$R(q) = I(X;Y). \tag{1.9}$$

The *information rate distortion function* is the minimum achievable rate over all conditional distributions $q(y|x)$ that satisfy the expected distortion constraint $\mathrm{E}[d(X,Y)] \leq D$:

$$R^{(I)}(D) = \min_{q(y|x):\int_{X,Y} f(x)q(y|x)(x-y)^2 \leq D} I(X;Y) \tag{1.10}$$

It can be shown that the inverse function $D^{(I)}(R)$ is well defined and could also be obtained by minimizing the distortion over all conditional distributions $q(y|x)$ that satisfy a rate constraint. We prefer the function $D^{(I)}(R)$ because it is related to the functions (1.6, 1.7) defined for quantization.

First of all, we have

$$D^{(I)}(R) \leq \bar{\delta}(R) \leq \delta_k(R), \tag{1.11}$$

[2]Please note that from here on, we will restrict ourselves to scalar random variables, since this thesis is only concerned with memoryless scalar sources. Also, we will often drop the indices on probability densities such as $q_{Y|X}(y|x)$ and just write $q(y|x)$ when there is no danger of confusion.

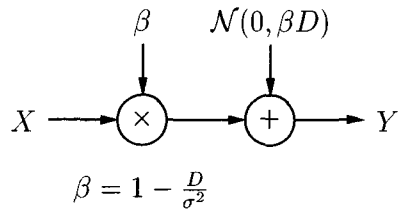

Figure 1.2: Forward test channel for the Gerrish-Schultheiss upper bound.

since the set of stochastic maps $q(y|x)$ includes every deterministic map $Q(x)$. In fact, the lossy source coding theorem first proved by Shannon [44, 45] implies that $D^{(I)}(R) = \bar{\delta}(R)$. Therefore from here on we will speak of the *distortion rate* function (drf) and denote it by $D(R)$. The adjective "operational" will only be used in conjunction with (classes of) quantizers.

Throughout the thesis we will use two classic bounds on $R(D)$. The *Gaussian upper bound* for a source of variance σ^2 is

$$R(D) \leq \frac{1}{2} \log \frac{\sigma^2}{D}, \tag{1.12}$$

and the *Shannon lower bound* (SLB) for a source with differential entropy $H(X)$ is

$$R(D) \geq H(X) - \frac{1}{2} \log(2\pi e D). \tag{1.13}$$

Most often we will actually use the corresponding bounds on the inverse function $D(R)$, that is

$$\frac{\sigma^2}{2\pi e} \exp\left(2H(\tfrac{X}{\sigma}) - 2R\right) \leq D(R) \leq \sigma^2 \exp(-2R). \tag{1.14}$$

For a large class of sources, the Shannon lower bound is asymptotically tight for $D \to 0$, respectively $R \to \infty$. In contrast, the Gaussian upper bound is very loose for the highly non-Gaussian, sparse source densities that are of interest in this thesis.

Some of the main results presented in the following chapters are upper bounds on $D(R)$, which we want to compare with previously known bounds. The only other bound that made it into textbooks, besides the Gaussian upper bound, is due to Gerrish and Schultheiss [18]:

$$R(D) \leq H(Y) - \frac{1}{2} \ln(2\pi e \beta D), \tag{1.15}$$

with $\beta = 1 - \frac{D}{\sigma^2}$ and the marginal density

$$q(y) = \int p(x) \frac{1}{\sqrt{2\pi\beta D}} \exp \frac{-(y - \beta x)^2}{2\beta D} \, \mathrm{d}x. \tag{1.16}$$

The bound can be easily derived using the test channel shown in Figure 1.2. If $H(Y)$ is upper bounded by the entropy of a Gaussian with the variance of Y, (1.15) reduces to the Gaussian upper bound. A disadvantage of the bound (1.15) is that it usually requires a double numeric integration to compute $H(Y)$, since the marginal $q(y)$ can rarely be obtained in closed form (as usual the Gaussian density is an exception).

Finally, we remark that one would expect that the conditional distribution $q^*(y|x)$ that achieves the minimum rate[3] in (1.10) is continuous if the source pdf $f(x)$ is also continuous. Astonishingly enough, this is usually not the case! At low rates, the support of Y is in most cases discrete, i.e. $q^*(y|x)$ exists only for $y \in \{y_1, y_2, \dots\}$. This phenomenon has been first described by Fix [16, 17] for sources with finite support. It was later independently rediscovered by Rose [38] who looked at the infinite support case. Rose also proved that the support set is continuous at sub-critical distortion (when the SLB is tight) and that it becomes discrete at supercritical distortion (SLB is not tight). Thus there is an anomaly in the computation of the rate distortion function precisely at low rates, which is the region that interests us most.

1.2 Outline of Thesis

We give an outline of the remaining chapters of this thesis:

- In Chapter 2 we introduce oligoquantization based on scalar thresholding and study the Gaussian case in some detail. Then we derive the main result, an upper bound on operational distortion rate for thresholding oligoquantization. We also provide tightened and loosened versions of the bound. The Chapter ends with an example where we compare the bound to actual quantizer performance.
- The third Chapter is concerned with transform coding, in particular we study several models for transform coefficients. Also, a tighter high-rate bound is derived and used in a new definition of coding gain.
- Chapter 4 contains two main results: the upside is a method to compute sample-based bounds without estimating a density. The downside is a theorem that shows that the principle of oligoquantization does not carry over well to high dimensions.
- Finally, Chapter 5 starts with a short summary of the thesis and then presents some concluding remarks.

1.2.1 Acronyms

Unfortunately also this author can't escape the damnation of communications research: the acronym. Table 1.1 is an incomplete list of those used throughout this thesis, with a reference to the defining section, if available.

[3] In general, there might actually be more than one such distribution.

Acronym	Meaning	Section
BMS	Binary memoryless source	
DMS	Discrete memoryless source	
drf	Distortion rate function	1.1
GG	Generalized Gaussian	2.5
GM	Gaussian Mixture	3.4
MCQ	Magnitude classifying quantization	3.5
MSE	Mean squared error	
pdf	Probability density function	
pmf	Probability mass function	
rdf	Rate distortion function	1.1
rv	Random variable	
SLB	Shannon lower bound	1.1
TOQ	Thresholding oligoquantization	2.1

Table 1.1: Some acronyms used in this thesis

Chapter 2

Bounding Low-Rate Distortion

In the introduction we have shown that a simple threshold-based oligoquantization method is a very efficient way to encode wavelet coefficients. The core of this chapter (and this thesis) will be an upper bound on operational distortion rate for thresholding oligoquantization (TOQ). After defining TOQ, we first study the behavior of thresholded Gaussian random variables. Then we derive the low-rate bound and a tightened version of it. In contrast, by loosening the bound we will gain more insight in the tradeoffs involved in TOQ. Finally, we present a simple example which allows us to compare the bounds with the distortion achieved by actual quantizers.

A Remark on Notation Even though the bit is with us since the beginnings of information theory, for the purpose of deriving analytical expressions it is much simpler to work with natural logarithms. Therefore in this thesis all *equations* involving entropies, rates, etc ... will be expressed in *nats*, and all logarithms are to the base e. Of course the standard disclaimer "unless otherwise noted" applies. To make things slightly confusing, all *plots* will use *bits* as a concession to all those who are eyeing for slopes of -6 dB/bit.

Another minor change from canonical notation concerns entropy: $h(\cdot)$ always stands for the binary entropy function $h(p) = -p \ln p - (1-p)\ln(1-p)$ (measured in nats), whereas $H(X)$ might stand for discrete or differential entropy, depending on the alphabet of X. This convention is motivated by some cases where both a binary entropy function and a differential entropy appear in the same expression.

2.1 Thresholding Oligoquantization

To keep things simple, most of this thesis considers only memoryless sources that emit a sequence of i.i.d. zero mean random variables with a *symmetric* probability density function (pdf) $f_X(x)$, that is

$$f_X(-x) = f_X(x).$$

This simplifies many computations, since the second moment of a set with zero mean will be equal to its variance:

$$\mathrm{E}[X|X \in \mathcal{S}] = 0 \iff \mathrm{E}[X^2|X \in \mathcal{S}] = \mathrm{Var}(X|X \in \mathcal{S}).$$

In particular, this is the case for a symmetric pdf when the set $\mathcal{S}$ is symmetric about the origin. Therefore we will often use the term variance to designate the second moment and vice-versa. This convention is used throughout the thesis, up to Section 4.1, where we will deal with arbitrary densities. As it turns out, many results carry over directly thanks to the fact that the second moment is an upper bound to the variance.

However, for now we will concentrate on symmetric densities, which is not a very severe restriction if we want to use our results to analyze low-rate lossy transform coding systems. For one, a main goal of using a transform is to have approximately independent coefficients at the output. In general these will not be identically distributed (think of the KLT), but mixing them into a single distribution will always provide an upper bound to the rate distortion performance of more elaborate vector coding schemes. Finally, the restriction to symmetric densities is minor too, because a transform that produces a skewed coefficient density is intuitively mismatched to the source. In fact most i.i.d. coefficient density models proposed in the literature are symmetric, e.g. Laplacian or generalized Gaussian.

Definition 2.1 *Thresholding oligoquantization (TOQ) of a sequence of i.i.d. random variables* $[X_i]_{i=1,\dots,\infty}$ *is a two step encoding process:*

1. *Each sample is classified: samples with magnitude above a threshold* T *are called* significant[1] *and assembled into a new sequence* $[V_j]$*, which might be empty. The other —* insignificant *— samples are discarded. The density of the thresholded random variable is*

$$\bar{f}^{(T)}(v) = \begin{cases} 0, & |v| < T \\ \frac{1}{\mu(T)} f(v), & |v| \geq T \end{cases} \tag{2.1}$$

Since the pdf $f(x)$ *is symmetric, the normalization constant* $\mu(T)$*, i.e. the probability of a significant sample, will be*

$$\mu(T) = \mathrm{E}[U] = \Pr\{|X| \geq T\} = 2\int_T^\infty f(x)\,dx. \tag{2.2}$$

A binary sequence

$$[U_i] = [1_{|X_i| \geq T}] \tag{2.3}$$

called significance map *records the positions of the significant samples and will be sent to the decoder as side information using rate* $h(\mu(T))$ *per sample (*h *is the binary entropy function).*

[1] The term *significant* is borrowed from image compression, where similar quantization methods are very popular.

2. *The sequence of* thresholded *random variables* $[V_j]$ *is then quantized with a* k*-dimensional quantizer matched to their density, yielding the quantized sequence* $[\hat{V}_j]$*. The quantizer indices are entropy coded and sent to the decoder.*

The decoder can reassemble the original sequence thanks to the significance map: the insignificant samples are mapped to the centroid $\mathrm{E}[X\,|\,|X| < T] = 0$*, while the significant samples are reconstructed from the quantized sequence* $[\hat{V}_j]$.

Often a scalar quantizer ($k = 1$) is used for the significant samples. Then the whole process can be described as a single scalar quantizer with a prescribed zero bin $[-T, T]$. Such scalar quantizers are also known as deadzone quantizers; if the bins outside the zero bin have the same width, we get a particular case of the uniform threshold quantizer [15].

The following proposition shows that two step TOQ with scalar quantization has the same asymptotic rate per sample (as the number of samples approaches infinity) as a single step quantizer with prescribed zero bin. The same threshold value T and the same quantization bins for the significant samples are used for both codes, therefore they yield the same distortion.

Proposition 2.1 *Let* $Y \in \{1, 2, \ldots, L\}$ *be the scalar quantization index of the* significant *values* $\hat{V}_j$ *in the two step code. Then the asymptotic rate per sample is* $R_{(2)} = h(\mu) + \mu H(Y)$*. Let* $\tilde{Y} \in \{0, 1, \ldots, L\}$ *be the scalar quantization index of the single step quantizer. The rate for this code is* $R_{(1)} = H(\tilde{Y}) = R_{(2)}$.

Proof: The significance map U_i can be coded with $h(\mu)$ bits per sample. For every $U_i = 1$ we have to encode a sample of the random variable Y, thus

$$R_{(2)} = h(\mu) + \mathrm{E}[U]H(Y) = h(\mu) + \mu H(Y).$$

If Y is distributed with $\Pr\{Y = i\} = p_i$, $i = 1, \ldots, L$, then $\tilde{Y}$ will have the distribution $\Pr\{\tilde{Y} = 0\} = 1 - \mu$, $\Pr\{\tilde{Y} = i\} = \mu p_i$, $i = 1, \ldots, L$. Therefore we have

$$\begin{aligned} R_{(1)} = H(\tilde{Y}) &= -(1-\mu)\log_2(1-\mu) - \sum_{i=1}^{L} \mu p_i \log_2(\mu p_i) \\ &= -(1-\mu)\log_2(1-\mu) - \mu\log_2(\mu) - \sum_{i=1}^{L} p_i \log_2(p_i) = R_{(2)} \end{aligned}$$

□

This means that the single step quantizer and two step TOQ are functionally equivalent, there is only an implementation difference.

Example 2.1 We consider two and four level Lloyd-Max quantization [29, 35] of thresholded Gaussian random variables. Figure 2.1 shows the operational distortion

rate curves obtained by using the threshold as a parameter. The two curves start at the respective (R, D) points for Lloyd-Max quantization ($T = 0$) and end in $(R, D) = (0, 1)$ for $T \to \infty$. They show that by varying the threshold one can achieve a whole set of (R, D) points which are actually below simple time sharing of the original Lloyd-Max solutions. Also shown are the Gaussian $D(R)$ as a lower bound, the Gish-Pierce asymptote for high-rate scalar quantization [21] and the operational $D(R)$ of entropy coded uniform quantization.

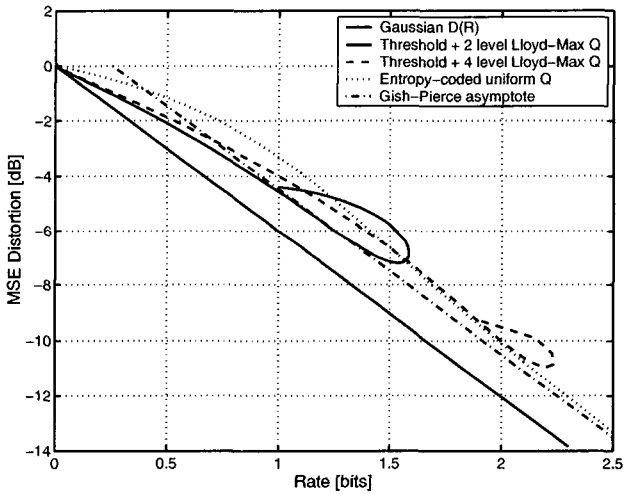

Figure 2.1: Thresholding oligoquantization of a Gaussian random variable.

2.2 R/D Analysis of Thresholded Gaussian

In order to derive rate distortion bounds for TOQ with arbitrary ("optimal") choice of threshold, we first characterize the R/D behavior of thresholded random variables. We restrict our analysis to the zero mean Gaussian case.

Upper bound We upper bound $D(R)$ of the thresholded Gaussian by the distortion rate function of a Gaussian with the same variance (1.12). From the Gaussian density, $f(x) = (\sqrt{2\pi}\sigma)^{-1} \exp(-x^2/2\sigma^2)$, we compute the pdf of the thresholded variable according to (2.1):

$$\bar{f}^{(T)}(v) = \begin{cases} 0, & |v| < T \\ \frac{1}{\mu(T)} \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}, & |v| \geq T \end{cases} \tag{2.4}$$

with normalization constant

$$\mu(T) = \operatorname{erfc}\left(\frac{\sqrt{2}}{2}\frac{T}{\sigma}\right). \tag{2.5}$$

The variance is $\mathrm{E}_{\tilde{f}}\, V^2 = A(T)/\mu(T)$ with

$$A(T) = 2\int_T^\infty x^2 f(x)\,\mathrm{d}x = \sigma^2\left(\mu(T) + \sqrt{\frac{2}{\pi}}\frac{T}{\sigma}e^{-\frac{T^2}{2\sigma^2}}\right). \tag{2.6}$$

Inserting this into the distortion rate formula for the Gaussian provides the desired upper bound:

$$D(R) \leq \tfrac{A(T)}{\mu(T)}e^{-2R}. \tag{2.7}$$

Shannon lower bound First, we compute the differential entropy of a thresholded Gaussian:

$$H(V) = \tfrac{A(T)}{2\mu(T)\sigma^2} + \tfrac{1}{2}\ln(2\pi\sigma^2) + \ln(\mu(T))$$

The Shannon lower bound follows from (1.13) as

$$\begin{aligned} R_{SLB}(D) &= H(V) - \tfrac{1}{2}\ln(2\pi eD) \\ &= \tfrac{A(T)-\mu(T)\sigma^2}{2\mu(T)\sigma^2} + \ln(\mu(T)) + \tfrac{1}{2}\ln\left(\tfrac{\sigma^2}{D}\right). \end{aligned} \tag{2.8}$$

Figure 2.2 shows the upper and lower bounds and the empirical $D(R)$ computed with Blahut's algorithm [4] for two values of the normalized threshold T/σ.[2] For larger thresholds, a distinct knee appears at $R = 1$. This is in accordance with the intuition that the first bit should give a large reduction in distortion for large threshold values. An interesting question is how close simple two level (one bit) quantization comes to the Shannon lower bound:

Proposition 2.2 *Let D_2 be the distortion for two level Lloyd-Max quantization of a thresholded Gaussian with pdf given by (2.4), and let D_{SLB} be the solution of $R_{SLB}(D) = \ln 2$. We have*

$$\lim_{T\to\infty}\frac{D_2}{D_{SLB}} = \frac{2\pi}{e}. \tag{2.9}$$

Thus the asymptotic distortion for two level quantization is about 2.31 times the lower bound, or 3.64 dB above it.

[2]Note that there is double normalization in Figure 2.2: first, the threshold t is taken relative to the standard deviation σ of the original rv, and after, the variance $\hat{\sigma}$ of the thresholded rv is normalized to one.

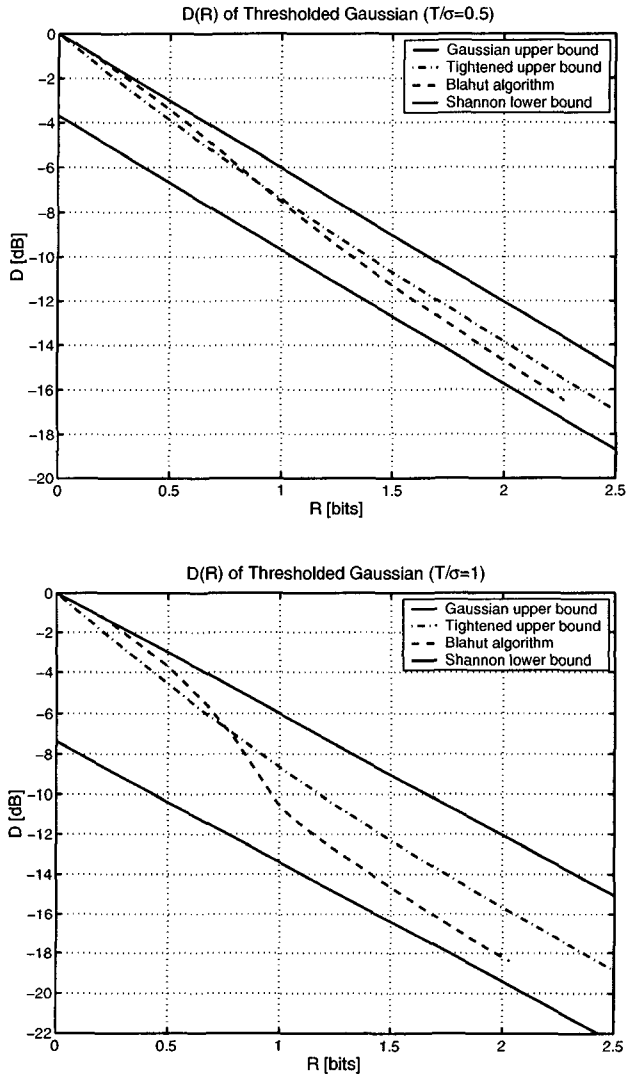

Figure 2.2: $D(R)$ behavior of thresholded Gaussian with $T/\sigma = 0.5$ (top) and $T/\sigma = 1$ (bottom): Upper and lower bounds and empirical $D(R)$ computed with Blahut's algorithm.

Proof: We use the substitutions $z = T/\sigma$, $a(z) = \frac{A(z\sigma)}{\mu(z\sigma)\sigma^2}$ and the relation $\mu(T) = \frac{\sqrt{2}ze^{-z^2/2}}{\sqrt{\pi}(a(z)-1)}$ to obtain

$$\begin{aligned} D_2 &= \left[a(z) - \left(\frac{a(z)-1}{z}\right)^2\right]\sigma^2 \\ D_{SLB} &= \frac{1}{2\pi}e^{a(z)-1-z^2}\frac{z^2}{(a(z)-1)^2}\sigma^2 \end{aligned}$$

and therefore

$$\lim_{T\to\infty}\frac{D_2}{D_{SLB}} = \lim_{z\to\infty}\left[\frac{2\pi}{e^{a(z)-1-z^2}}\right]\cdot\left[\frac{(a(z)-1)^2}{z^2}\left(a(z) - \frac{(a(z)-1)^2}{z^2}\right)\right]$$

Inserting $a(z) - 1 = \frac{\sqrt{2}ze^{-z^2/2}}{\sqrt{\pi}\,\mathrm{erfc}(z/\sqrt{2})}$ and applying Bernoulli-de l'Hospital's rule repeatedly shows that the first term converges to $\frac{2\pi}{e}$ and the second to one. □

This result should be compared to the asymptotic (high-rate) performance achievable with scalar quantization. Gish and Pierce[21] found that high-rate uniform quantizers are optimal and yield a distortion that is $\pi e/6 \approx 1.42$ times the Shannon lower bound. Also, one bit quantization of a Gaussian has a distortion of $4(\pi-2)/\pi \approx 1.45$ times the SLB. We conclude that even low-rate (1 bit) scalar quantization of the thresholded Gaussian is quite efficient. In fact, since the gap between the upper bound (2.7) and the lower bound (2.8) widens for growing T, we see that scalar quantization of the thresholded samples has a distortion that is asymptotically (for $T \to \infty$) vanishing compared to the upper bound.

2.2.1 A Tighter Upper Bound

As just mentioned, the Gaussian upper bound is very loose for large thresholds. To get a tighter bound, we make the following thought experiment: let us use a Gaussian codebook to encode a Gaussian random variable X with average distortion $D(R)$. Now assume that we get the *free* side information $U = 1_{|X|\geq T}$ which we can use to improve the estimate $\hat{X}$. Write the average distortion as

$$\begin{aligned} D(R) &= \Pr\{U=0\}\,\mathrm{E}\,d(X,\hat{X}|U=0) + \Pr\{U=1\}\,\mathrm{E}\,d(X,\hat{X}|U=1) \\ &= (1-\mu(T))\,D_0 + \mu(T)\,D_1, \end{aligned} \tag{2.10}$$

with $D_i = \mathrm{E}\,d(X,\hat{X}|U=i)$. It is evident that D_1 is the thresholded Gaussian $D(R)$ that we want to bound. Since we know the random codebook distribution needed to achieve $D(R)$, we can easily compute D_1. The goal is now to find a function $\phi(\hat{x})$ that minimizes $\widetilde{D}_1 = Ed(X,\phi(\hat{X})|U=1)$. By the calculus of variations we find

$$\phi^*(\hat{x}) = \hat{x} + \sqrt{\frac{2D}{\pi}}\,\frac{e^{\frac{(\hat{x}+T)^2}{2D}} - e^{\frac{(\hat{x}-T)^2}{2D}}}{\mathrm{erf}\left(\frac{\hat{x}+T}{\sqrt{2D}}\right) - \mathrm{erf}\left(\frac{\hat{x}-T}{\sqrt{2D}}\right) - 2}. \tag{2.11}$$

Unfortunately this expression is hard to integrate symbolically, but its graph served as an inspiration for a simpler function,

$$\phi(\hat{x}) = \hat{x} + a\hat{x}e^{-b\hat{x}^2}, \tag{2.12}$$

that is easier to handle and gives quite tight bounds when optimized over the parameters a, b. Actually, setting $a = b = 0$ already gives a tighter bound than (2.7), namely

$$D_1(R) \leq \left(1 + \frac{\sqrt{2}\frac{T}{\sigma}e^{-\frac{T^2}{2\sigma^2}}}{\sqrt{\pi}\mu(T)}e^{-2R}\right)\sigma^2 e^{-2R} \tag{2.13}$$

A plot of this bound is included in Figure 2.2. Rather disturbingly, the bound crosses the $D(R)$ curve computed with Blahut's algorithm (BA). The culprit is the discretization of the density needed for the BA, which introduces some error. Additionally, normalizing by the variance of the discretized source introduces even more error.

2.3 Upper Bound on Distortion Rate

Now we set out to upper bound the operational $D(R)$ performance of TOQ, which is implicitly also an upper bound to the distortion rate function (drf) of the source being quantized. This is the key result of this thesis, since it will allow us to characterize the low-rate $D(R)$ behavior of "highly non-Gaussian" sources.

Thresholding oligoquantization is a two step process, consisting of thresholding followed by quantization of the significant samples. Ideally one would use a variable-dimension vector quantizer, since the number of samples varies from block to block (assuming data is processed in blocks). In practical schemes, scalar quantization is often preferred thanks to its simplicity and speed. And at low rates it performs quite well, as illustrated by Example 2.1.

In general we can never achieve a rate distortion optimal encoding if we use two step TOQ for lossy encoding of a block of i.i.d. random variables. Except in "constructed" cases, the thresholding implies a suboptimal code even if it is followed by an optimal quantizer for the significant samples. An optimal two step encoder is composed of an ideal lossless encoder for the significance indicator U, and an R/D optimal encoder for the values V of the significant samples. Therefore we first characterize the R/D behavior of thresholded random variables.

The $D(R)$ function of a single thresholded random variable V can be upper bounded with the distortion rate function of a Gaussian with the same variance. The unnormalized variance is

$$A(T) = 2\int_T^\infty f(x)x^2\,\mathrm{d}x \tag{2.14}$$

where $f(x)$ is the pdf of the non-thresholded rv and $A(0) = \sigma^2$ its variance. After normalization with (2.2) the upper bound becomes

$$D_V(R_V) \leq \mathrm{E}_{\tilde{f}}\, V^2 e^{-2R_V} = \tfrac{A(T)}{\mu(T)} e^{-2R_V} \tag{2.15}$$

We will not use tighter upper bounds such as the one presented in Section 2.2.1, since they would complicate if not inhibit the optimization steps that lead to the main result.

For each sample we need $H(U) = h(\mu(T))$ bits to indicate whether it is *significant*, i.e. above threshold. Given the total rate per sample R, the rate R_V available to code each significant sample is

$$R_V(T, R) = \frac{R - h(\mu(T))}{\mu(T)}. \tag{2.16}$$

Taking into account the distortion resulting from the uncoded sub-threshold samples, we can write the following bound on the drf of the source X:

$$D(R) \leq \mu(T) D_V(R_V(T, R)) + (1 - \mu(T))\, \mathrm{E}[X^2 \,|\, |X| < T]$$

or

$$D(R) \leq B(T, R) = A(T) \exp\left(-2\tfrac{R - h(\mu(T))}{\mu(T)}\right) + \sigma^2 - A(T). \tag{2.17}$$

The sought-after TOQ bound is obtained by optimizing over the threshold t, that is we search the rate $R^*(t_0)$ for which the bound $D(R^*(t_0)) \leq B(R^*(t_0))$ is tightest over all t in some neighborhood of t_0. This means that the bound will only be locally optimal (see remark after Theorem 2.3).

Direct minimization of $B(t, R)$ with respect to t is very difficult. Instead we sweep the threshold t from 0 to ∞ and compute candidate points of the convex hull of the resulting bounds $D(R) \leq B(t, R)$ in the (R, D) plane. To further simplify the task, we assume that two bounds $B(t, r)$ and $B(t + \Delta t, r)$ have a common tangent (equivalently: they intersect each other). This tangent can then be used to find candidate convex hull points. The resulting parametric upper bound will be called the *magnitude bound*, since the involved thresholding is only concerned with the magnitude of the samples.

Theorem 2.3 (Magnitude bound) *The operational distortion rate function for thresholding oligoquantization of a memoryless continuous random variable X with symmetric pdf $f(x)$ and variance σ^2 is upper bounded by*

$$D(R^*(t)) \leq A(t) \left[\exp\left(-2\tfrac{R^*(t) - h(\mu(t))}{\mu(t)}\right) - 1\right] + \sigma^2, \quad \forall\, t \geq 0 : \exists R^*(t) \tag{2.18}$$

where the rate $R^*(t)$ is given by (primes denote derivatives)

$$R^*(t) = h(\mu(t)) - \tfrac{1}{2}\mu(t) \left[2h'(\mu(t)) + \gamma(t) + W_{-1}\left(-\gamma(t) e^{-2h'(\mu(t)) - \gamma(t)}\right)\right] \tag{2.19}$$

with

$$\gamma(t) = \frac{\mu(t)}{A(t)} t^2 \tag{2.20}$$

The expression for $R^*(t)$ involves Lambert's W function, which solves $W(x)e^{W(x)} = x$. The subscript -1 indicates the second real branch of W, taking values on $[-1, -\infty[$. A real-valued solution for R^* exists only if the argument of the W function is larger than $-1/e$ (see remarks after proof).

Proof:

We take two curves of the family (2.17), say $B(t, r_0)$ and $B(t + \Delta t, r_1)$, and determine their common tangent by solving the following system of equations:

$$(\tfrac{\partial}{\partial r} B)(t, r_0) = (\tfrac{\partial}{\partial r} B)(t + \Delta t, r_1) = s \tag{2.21}$$

$$\frac{B(t + \Delta t, r_1) - B(t, r_0)}{r_1 - r_0} = s. \tag{2.22}$$

Using $(\frac{\partial}{\partial r} B)(t, r) = -2\frac{A(t)}{\mu(t)} \exp\left(-2\frac{r - h(\mu(t))}{\mu(t)}\right)$ we solve (2.21) for r_1:

$$r_1 = \tfrac{\mu(t+\Delta t)}{\mu(t)}[r_0 - h(\mu(t))] + h(\mu(t + \Delta t)) - \tfrac{1}{2}\mu(t + \Delta t) \ln \tfrac{A(t)\mu(t+\Delta t)}{A(t+\Delta t)\mu(t)} \tag{2.23}$$

and start inserting this solution into (2.22):

$$\begin{aligned} s &= \frac{B(t + \Delta t, r_1) - B(t, r_0)}{r_1 - r_0} \\ &= \frac{A(t + \Delta t) e^{-2\frac{r_0 - h(\mu(t))}{\mu(t)} + \ln \frac{A(t)\mu(t+\Delta t)}{A(t+\Delta t)\mu(t)}} - A(t) e^{-2\frac{r_0 - h(\mu(t))}{\mu(t)}}}{r_1 - r_0} \\ &\quad - \frac{[A(t + \Delta t) - A(t)]}{r_1 - r_0} \\ &= \frac{[\mu(t + \Delta t) - \mu(t)]\frac{A(t)}{\mu(t)} e^{-2\frac{r_0 - h(\mu(t))}{\mu(t)}} - [A(t + \Delta t) - A(t)]}{r_1 - r_0} \\ &= (\tfrac{\partial}{\partial r} B)(t, r_0) = -2\frac{A(t)}{\mu(t)} e^{-2\frac{r_0 - h(\mu(t))}{\mu(t)}}, \end{aligned} \tag{2.24}$$

where the last equality is actually again (2.21). 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

$$\frac{A(t)}{\mu(t)} e^{-2\frac{r_0 - h(\mu(t))}{\mu(t)}} [\Delta\mu + 2(r_1 - r_0)] - \Delta A = 0. \tag{2.25}$$

Now we let

$$y = -2\frac{r_0}{\mu(t)}, \tag{2.26}$$

$$\alpha = \frac{A(t)}{\mu(t)} e^{2\frac{h(\mu(t))}{\mu(t)}}, \tag{2.27}$$

$$\begin{aligned} \beta &= \Delta\mu + 2(r_1 - r_0) - 2\tfrac{r_0}{\mu(t)}\Delta\mu \\ &= \Delta\mu - 2\tfrac{\mu(t+\Delta t)}{\mu(t)} h(\mu(t)) + 2h(\mu(t + \Delta t)) - \mu(t + \Delta t) \ln \tfrac{A(t)\mu(t+\Delta t)}{A(t+\Delta t)\mu(t)} \end{aligned} \tag{2.28}$$

so that (2.25) becomes

$$\alpha e^{y}(-\Delta\mu\, y + \beta) - \Delta A = 0. \tag{2.29}$$

At this point we have to make an additional assumption about the pdf $f(x)$, namely that it is not constantly equal to zero over the interval $[t, t + \Delta t)$ (as long as this interval is contained in the support of f). Then for any $\Delta t > 0$ we have $\Delta\mu < 0$ (and $\Delta A < 0$) by definition.[3] Therefore we can divide (2.29) by $-\alpha\Delta\mu$ and get

$$e^{y}(y - \tfrac{\beta}{\Delta\mu}) + \tfrac{\Delta A}{\alpha\Delta\mu} = 0. \tag{2.30}$$

Now we can finally use the Lambert W function of Maple fame to find the solution(s)

$$y = \tfrac{\beta}{\Delta\mu} + W\left(-\tfrac{\Delta A}{\alpha\Delta\mu} e^{-\frac{\beta}{\Delta\mu}}\right). \tag{2.31}$$

Using the defining equation $W(x)e^{W(x)} = x$ it is easy to show that (2.31) actually solves (2.30). Before we go on and take limits, let's resolve the question on what branch of W we have to use. From (2.26) 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 (2.26) and (2.17)), 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 (2.31) 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 the argument of W:

$$\begin{aligned}\lim_{\Delta t\to 0} \tfrac{\beta}{\Delta\mu} &= \lim_{\Delta t\to 0} 1 + 2\tfrac{\mu(t)h(\mu(t+\Delta t)) - \mu(t+\Delta t)h(\mu(t)) + \mu(t)h(\mu(t)) - \mu(t)h(\mu(t))}{\mu(t)[\mu(t+\Delta t) - \mu(t)]} \\ &\quad + \tfrac{\mu(t+\Delta t)\,\Delta t}{\mu(t+\Delta t) - \mu(t)} \left[\tfrac{\ln A(t+\Delta t) - \ln A(t)}{\Delta t} - \tfrac{\ln \mu(t+\Delta t) - \ln \mu(t)}{\Delta t}\right] \\ &= 1 + 2h'(\mu(t)) - 2\tfrac{h(\mu(t))}{\mu(t)} + \tfrac{\mu(t)A'(t)}{\mu'(t)A(t)} - 1, \end{aligned} \tag{2.32}$$

$$\begin{aligned}\lim_{\Delta t\to 0} \tfrac{\Delta A}{\alpha\Delta\mu} &= \lim_{\Delta t\to 0} \tfrac{[A(t+\Delta t) - A(t)]\mu(t)}{[\mu(t+\Delta t) - \mu(t)]A(t)} e^{-2h(\mu(t))/\mu(t)} \\ &= \tfrac{A'(t)\mu(t)}{A(t)\mu'(t)} e^{-2h(\mu(t))/\mu(t)}. \end{aligned} \tag{2.33}$$

By the definitions of μ and A we have $\mu'(t) = -2f(t)$ and $A'(t) = -2f(t)t^2$, hence $\frac{\mu(t)A'(t)}{\mu'(t)A(t)} = \frac{\mu(t)}{A(t)}t^2 = \gamma(t)$, as defined in (2.20). Inserting (2.32, 2.33) into (2.31) gives

$$y = 2h'(\mu(t)) - 2\tfrac{h(\mu(t))}{\mu(t)} + \gamma(t) + W\left(-\gamma(t)e^{-2h'(\mu(t)) - \gamma(t)}\right) \tag{2.34}$$

[3] If this assumption does not hold, we also have $\beta = 0$ and thus (2.26) is always satisfied. This case corresponds to a random variable with holes in its support, that is a mixture of two (or more) random variables with non-overlapping supports (ranges). Picking the "critical" t actually separates the mixture components in two groups.

and after inserting this into (2.26) and solving for r_0 we get (2.19). □

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 is the case when increasing t produces a bound $B(t+\Delta t, r)$ that for all r is larger than $B(t,r)$. Using a Taylor approximation

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

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

$$\min_r (\tfrac{\partial}{\partial t}B)(t,r) > 0. \tag{2.35}$$

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

$$(\tfrac{\partial^2}{\partial t \partial r}B)(t,r) = 2e^{-2\frac{r-h(\mu(t))}{\mu(t)}} \cdot \left[\frac{-\mu^2(t)A'(t)+2\mu'(t)A(t)\left(\frac{1}{2}\mu(t)-r+h(\mu(t))-\mu(t)h'(\mu(t))\right)}{\mu^3(t)}\right]. \tag{2.36}$$

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

$$r^* = \tfrac{1}{2}\mu(t)\left[1-\gamma(t)-2h'(\mu(t))+2\tfrac{h(\mu(t))}{\mu(t)}\right]. \tag{2.37}$$

To verify that it is indeed a (unique, thus global) minimum:

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

Now we insert (2.37) into (2.35) to characterize those t that satisfy

$$\begin{aligned}(\tfrac{\partial}{\partial t}B)(t,r^*) &= \left[A'(t)+2\tfrac{A(t)\mu'(t)}{\mu(t)}\left(h'(\mu(t))+\tfrac{r^*-h(\mu(t))}{\mu(t)}\right)\right]e^{-2\frac{r^*-h(\mu(t))}{\mu(t)}} - A'(t)\\ &= \tfrac{A(t)\mu'(t)}{\mu(t)}e^{\gamma(t)-1+2h'(\mu(t))} - A'(t).\end{aligned} \tag{2.38}$$

If we divide (2.38) by $A'(t) = -2f(t)t^2 < 0$ (thus flipping the inequality) we see that (2.35) is equivalent to $e^{-1}e^{\gamma(t)+2h'(\mu(t))} < \gamma(t)$ and after another sign flip we get

$$-\tfrac{1}{e} > -\gamma(t)e^{-\gamma(t)-2h'(\mu(t))}. \tag{2.39}$$

This is a reassuring result: condition (2.39), and thus (2.35), is true exactly iff the argument of the W function in (2.19) is less than or equal to $-1/e$, 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 $B(t,r)$ might be

below all others for a threshold t satisfying (2.35). In all examples we have studied, it happened to be the trivial bound $B(0, r)$, if at all. For a Gaussian source, obviously no upper bound can be better than $B(0, r)$, while for Laplacian sources all thresholds above a critical value yield slight improvements. For pdf's that are even more peaked around zero, such as those studied in the next chapter, 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 position rate $h(\mu(t))$. 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 (i.e. for high thresholds), a rate region for which few bounds are known.

An example of the bound is plotted in Figure 2.3, together with an improved version that will be presented next.

2.4 Tightening the Bound

The approach to tighten the bound is based on quantization in the sense that we split the set of significant samples according to their sign. Thus we get a total of three classes: "0", "+" and "−". As before, the insignificant samples (the "0" class) will be mapped to zero. By symmetry of $f(x)$ the rate R_V will be divided evenly among "+" and "−" samples; the bound can be derived with the same techniques as for Theorem 2.3. However, besides the probability mass and the second moment, the first moment of the significant samples must also be known, or else the bound will not be very tight.

The magnitude bound is based on classifying the samples into significant and insignificant ones. If the threshold is fairly large, the pdf of the significant samples will have a sizeable gap around zero, and their variance will be much larger than twice the variance of the positive (or negative) significant samples alone. Thus we can expect a tighter bound if we further classify the significant samples according to their sign, even though the side information rate increases. Since the pdf $f(x)$ is assumed to be symmetric, we need one bit more per significant sample. The average rate becomes

$$R = h(\mu(T)) + \mu(T)(\ln 2 + R_V),$$

(the $\ln 2$ term is the sign *bit* in *nats*) and the corresponding distortion bound is

$$\begin{aligned} D(R) \leq\ & \mu(T)\,\mathrm{Var}(X|X \geq T)e^{-2\frac{R-h(\mu(T))-\mu(T)\ln 2}{\mu(T)}} \\ & + (1-\mu(T))\,\mathrm{Var}(X||X| < T). \end{aligned} \tag{2.40}$$

The insignificant samples are quantized to zero as before, while the significant samples are now one-sided random variables with variance $\mathrm{Var}(X|X \geq T)$, since we know their sign. One disadvantage of this approach is that besides $\mu(t)$ and $A(t)$, we also

need the first moment (the *centroid*)

$$\xi(t) = \mathrm{E}[X|X \geq t] = \frac{\int_t^\infty f(x)x\,dx}{\mu(t)/2} \tag{2.41}$$

to compute that variance. This is the price to be paid for a (sometimes) tighter bound.

Theorem 2.4 (Magnitude plus Sign Bound) *The operational distortion rate function for thresholding oligoquantization of a memoryless continuous random variable X with symmetric pdf $f(x)$ and variance σ^2 is upper bounded by*

$$\begin{aligned} D(R^*_{ms}(t)) &\leq B_{ms}(R^*_{ms}(t)) \\ &= 4[A(t) - \mu(t)\xi^2(t)]e^{-2\frac{R^*_{ms}(t)-h(\mu(t))}{\mu(t)}} + \sigma^2 - A(t), \end{aligned} \tag{2.42}$$

where the rate $R^*_{ms}(t)$ is given by

$$\begin{aligned} R^*_{ms}(t) &= h(\mu(t)) - \mu(t)h'(\mu(t)) \\ &\quad - \tfrac{\mu(t)}{2}\left[\gamma_1(t) + W_{-1}\left(-\gamma_2(t)e^{-\gamma_1(t)-2h'(\mu(t))}\right)\right] \end{aligned} \tag{2.43}$$

with

$$\gamma_1(t) = \frac{(\xi(t)-t)^2}{\frac{A(t)}{\mu(t)} - \xi^2(t)} \quad \textit{and} \quad \gamma_2(t) = \frac{\frac{1}{4}t^2}{\frac{A(t)}{\mu(t)} - \xi^2(t)}. \tag{2.44}$$

Proof: The proof works along the same lines as the one of Theorem 2.3. Instead of (2.23) the solution for r_1 is now

$$\begin{aligned} r_1 &= \tfrac{\mu(t+\Delta t)}{\mu(t)}[r_0 - h(\mu(t))] + h(\mu(t+\Delta t)) \\ &\quad - \tfrac{1}{2}\mu(t+\Delta t)\ln\tfrac{[A(t)-\mu(t)\xi^2(t)]\mu(t+\Delta t)}{[A(t+\Delta t)-\mu(t+\Delta t)\xi^2(t+\Delta t)]\mu(t)} \end{aligned} \tag{2.45}$$

Equation (2.24) becomes

$$\begin{aligned} s &= \frac{B_3(t+\Delta t, r_1) - B_3(t, r_0)}{r_1 - r_0} = \frac{\Delta\mu\left(\frac{A(t)}{\mu(t)} - \xi^2(t)\right)e^{2-2\frac{r_0-h(\mu(t))}{\mu(t)}} - \Delta A}{r_1 - r_0} \\ &= (\tfrac{\partial}{\partial r}B_3)(t, r_0) = -2\left(\frac{A(t)}{\mu(t)} - \xi^2(t)\right)e^{2-2\frac{r_0-h(\mu(t))}{\mu(t)}}, \end{aligned} \tag{2.46}$$

which after the substitutions (compare with (2.26-2.28))

$$y = -2\tfrac{r_0}{\mu(t)}, \tag{2.47}$$

$$\alpha = 4\left(\tfrac{A(t)}{\mu(t)} - \xi^2(t)\right)e^{2\frac{h(\mu(t))}{\mu(t)}}, \tag{2.48}$$

$$\begin{aligned} \beta &= \Delta\mu + 2(r_1 - r_0) - 2\tfrac{r_0}{\mu(t)}\Delta\mu \\ &= \Delta\mu - 2\tfrac{\mu(t+\Delta t)}{\mu(t)}h(\mu(t)) + 2h(\mu(t+\Delta t)) \\ &\quad - \mu(t+\Delta t)\ln\frac{\frac{A(t)}{\mu(t)} - \xi^2(t)}{\frac{A(t+\Delta t)}{\mu(t+\Delta t)} - \xi^2(t+\Delta t)} \end{aligned} \tag{2.49}$$

is again equivalent to

$$\alpha e^{y}(-\Delta\mu\, y + \beta) - \Delta A = 0. \tag{2.50}$$

Therefore we can use solution (2.31), but now inserting the following limits:

$$\lim_{\Delta t\to 0} \frac{\beta}{\Delta\mu} = 2h'(\mu(t)) - 2\frac{h(\mu(t))}{\mu(t)} + \frac{(\xi(t)-t)^2}{\frac{A(t)}{\mu(t)} - \xi^2(t)}, \tag{2.51}$$

$$\lim_{\Delta t\to 0} \frac{\Delta A}{\alpha\Delta\mu} = \frac{\frac{1}{4}t^2}{\frac{A(t)}{\mu(t)} - \xi^2(t)} e^{-2h(\mu(t))/\mu(t)}. \tag{2.52}$$

The theorem follows. □

For lazy people, or when $\xi(t)$ is not available, Theorem 2.4 can be weakened to

Corollary 2.5 (Weak Magnitude plus Sign Bound)

$$\begin{aligned} D(R^*_{wms}(t)) &\le B_{wms}(R^*_{wms}(t)) \\ &= 4[A(t) - \mu(t)t^2]e^{-2\frac{R^*_{wms}(t)-h(\mu(t))}{\mu(t)}} + \sigma^2 - A(t), \end{aligned} \tag{2.53}$$

with

$$R^*_{wms} = h(\mu(t)) - \mu(t)h'(\mu(t)) - \tfrac{\mu(t)}{2} W_{-1}\left(-\tfrac{\gamma(t)}{4(1-\gamma(t))} e^{-2h'(\mu(t))}\right) \tag{2.54}$$

Proof: $\xi(t) \ge t$ therefore $\mathrm{Var}(X|X \ge t) \le \frac{A(t)}{\mu(t)} - t^2$. □

Figure 2.3 shows the upper bounds and the Shannon lower bound for a Laplacian source with pdf $f(x) = \frac{\sqrt{2}}{2} e^{\sqrt{2}|x|}$, magnitude bound (2.18), magnitude+sign bound (2.42) and weakened M+S bound (2.53). Clearly the latter is weaker than the other two; only at rates below 0.01 bits it becomes slightly tighter. For more peaked densities this situation improves a bit, but in general Corollary 2.5 is rather useless when compared to the magnitude bound (Theorem 2.3). On the other hand, as announced the magnitude+sign bound is tightest at lower rates up to about 2 bits.

Even though the M+S bound beats the Gaussian upper bound by about 0.2 dB at 0.2 bits, these bounds are not overwhelming. The examples in the next section will show that they get more interesting for densities that are more peaked than the Laplacian. Here interesting means upper bounds that are substantially below the Gaussian upper bound, corresponding to sources that are efficiently quantized with a thresholding oligoquantization.

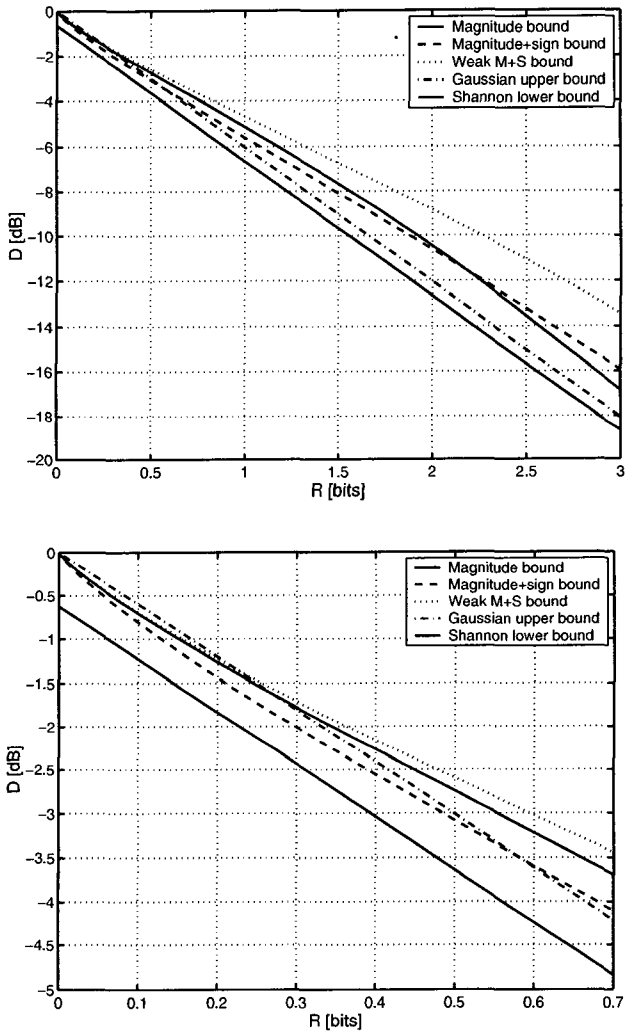

Figure 2.3: *Distortion-rate upper bounds for a Laplacian source (unit variance). Low-rate detail at bottom.*

2.5 Loosening the Bound

The primary objective of loosening the bound is to have simpler expressions to deal with. In turn these will give better insight into the "geometry" of low-rate TOQ. Since the bound (2.17) holds for all rates R and thresholds T, we can use an approximation to $R^*(T)$ without giving up the bounding property; the resulting bound will simply be less tight. In other words, any real-valued function $R(t)$ will yield an upper bound, but the goal is to approximate $R^*(t)$ at low rates. The most troublesome term in the expression for $R^*(t)$ is of course the Lambert W function, which makes it an ideal candidate for an approximation. From the paper by Corless et. al. [9] we get the following series expansion:

$$W_{-1}(z) = \ln(-z) - \ln(-\ln(-z)) - \sum_{k\geq 0} \ln(1 + p_k/v_k), \tag{2.55}$$

where $v_{n+1} = v_n + p_n$ and $p_{n+1} = -\ln(1+p_n/v_n)$, with starting values $v_0 = \ln(-z)$ and $p_0 = \ln(-\ln(-z))$, respectively. For $z \to 0-$ we have $p_0/v_0 \to 0$, thus the first two terms in (2.55) are a good approximation for $W_{-1}(z)$ if z is small.

Looking at (2.19) we get

$$W_{-1}\left(-\gamma(t)e^{-2h'(\mu(t))-\gamma(t)}\right) \approx \ln\gamma(t) - 2h'(\mu(t)) - \gamma(t) - \ln\left[-\ln\gamma(t) + 2h'(\mu(t)) + \gamma(t)\right]$$

for small $\mu(t)$ (since $0 < \gamma(t) \leq 1$) and finally

$$R^*(t) \approx R_1(t) = h(\mu(t)) + \frac{\mu(t)}{2}\left\{-\ln\gamma(t) + \ln\left[-\ln\gamma(t) + 2h'(\mu(t)) + \gamma(t)\right]\right\}. \tag{2.56}$$

Using this approximation in (2.18) yields the following weaker upper bound:

$$\begin{aligned} D(R_1(t)) &\leq \frac{A(t)\gamma(t)}{-\ln\gamma(t) + 2h'(\mu(t)) + \gamma(t)} + \sigma^2 - A(t) \\ &= \frac{\mu(t)\,t^2}{-\ln\gamma(t) - 2\ln\mu(t) + 2\ln(1-\mu(t)) + \gamma(t)} + \sigma^2 - A(t) \end{aligned} \tag{2.57}$$

Since $\lim_{z\to 0-} p_0/v_0 = 0$, the first term in (2.55) dominates the second term. The former can therefore be used to get an even simpler approximation:

$$R^*(t) \approx R_2(t) = h(\mu(t)) - \frac{\mu(t)}{2}\ln\gamma(t). \tag{2.58}$$

Inserting this into (2.18) gives

$$D(R_2(t)) \leq A(t)\gamma(t) + \sigma^2 - A(t) = \mu(t)\,t^2 + \sigma^2 - A(t) \tag{2.59}$$

with $R(t)$ given by the right hand side of (2.58). However the decay of $\ln(x)/x$ is very slow, so (2.59) will only be useful for vanishingly small μ, that is $R \to 0$. The above nicely displays the basic interaction between $\mu(t)$ and $A(t)$ in bounding distortion rate. It also allows one to upper bound the slope of $D(R)$ at $R = 0$.

Theorem 2.6 *Let $f(x)$ be a symmetric, finite variance pdf that satisfies the conditions*

(i) $\lim_{t\to\infty} f(t) = 0$ *and* $f'(t)$ *exists for* $t \to \infty$,

(ii) $\mu(t) > 0$ *(and* $A(t) > 0$*) for any finite* $t \geq 0$.

Then the slope[4] *of $D(R)$ at $R = 0$ satisfies*

$$D'(R)\Big|_{R=0} \leq \lambda_0 = -2\left(\lim_{t\to\infty} \frac{f(t)}{f'(t)}\right)^2. \tag{2.60}$$

Proof: First, a note about the conditions: the first is straightforward and the second simply guarantees that the support of $f(x)$ is the whole real line (if $supp(f)$ were say $[-t_1, t_1]$, one would use left-hand limits at t_1). Note also that by the dominated convergence theorem of real analysis [41] we have

$$\lim_{t\to\infty} \mu(t) = 0 \text{ and } \lim_{t\to\infty} A(t) = 0. \tag{2.61}$$

From the conditions we can conclude that

$$0 < \lim_{t\to\infty} \gamma(t) \leq 1, \tag{2.62}$$

which we will need later. Recalling the definition of γ, namely $\gamma(t) = \mu(t)\, t^2/A(t) = t^2/\mathrm{E}[X^2 | X \geq t]$, we see that condition (iii) guarantees that the limit in (2.62) is well defined. Then the right-hand inequality is satisfied by definition, while the left-hand one follows from $\mu(t)\, t^2 > 0$ and $A(t) < \sigma^2$ for $t > 0$.

Let $R(t) = h(\mu(t)) - \frac{\mu(t)}{2} \ln \gamma(t)$ as in (2.58), then $\lim_{t\to\infty} R(t) = 0$ by (2.61) and (2.62). Let $B(t) = A(t)\gamma(t) + \sigma^2 - A(t)$, the right-hand side of (2.59). Then we also have $\lim_{t\to\infty} B(t) = \sigma^2 = D(0)$ and since B is an upper bound,

$$D'(0) \leq \lim_{t\to\infty} \frac{B'(t)}{R'(t)}.$$

That is, at $R = 0$ the slope of $D(R)$ must be more negative than the slope of the upper bound (2.59). To compute this limit, we observe that $\mu'(t) = -2f(t)$ and

[4]Note to those who try to verify this on a graph: the rate is measured in *nats* ...

$A'(t) = -2f(t)\,t^2$ to obtain

$$\begin{aligned}\lim_{t\to\infty}\frac{B'(t)}{R'(t)} &= \lim_{t\to\infty}\frac{2\mu(t)t}{2f(t)[\ln\mu(t)-\ln(1-\mu(t))+\ln\gamma(t)+1-\gamma(t)]-\mu(t)/t}\\ &= \lim_{t\to\infty}\frac{\mu(t)\,t}{f(t)\ln\mu(t)} = \left(\lim_{t\to\infty}\frac{\mu(t)}{f(t)}\right)\left(\lim_{t\to\infty}\frac{t}{\ln\mu(t)}\right)\\ &= \left(\lim_{t\to\infty}\frac{-2f(t)}{f'(t)}\right)\left(\lim_{t\to\infty}\frac{\mu(t)}{-2f(t)}\right)\\ &= -2\left(\lim_{t\to\infty}\frac{f(t)}{f'(t)}\right)^2 = \lambda_0\end{aligned}$$

□

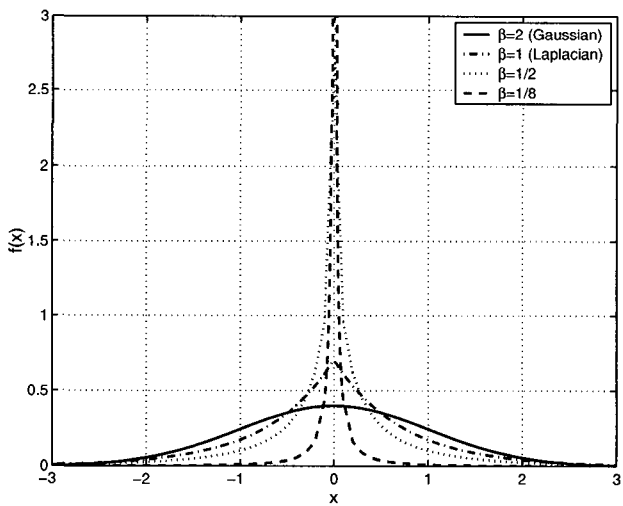

Figure 2.4: Generalized Gaussian pdf for several values of the shape parameter β (all with variance $\sigma^2 = 1$).

Example 2.2 Consider a random variable with a generalized Gaussian density

$$f(t) = \frac{\beta}{2\alpha\Gamma(\beta^{-1})}\exp\left(-\frac{|t|^\beta}{\alpha^\beta}\right)$$

with two parameters: $\alpha > 0$ and the *shape parameter* $\beta > 0$.

Remark: This notation is used e.g. in [31], whereas the more "canonical" notation is $f(x) = \frac{\nu b(\sigma,\nu)}{2\Gamma(\nu^{-1})} \exp(-b(\sigma,\nu)|x|^{\nu})$ with $b(\sigma,\nu) = \frac{1}{\sigma}\sqrt{\Gamma(3\nu^{-1})/\Gamma(\nu^{-1})}$ determined by the shape parameter ν and the standard deviation σ. The correspondence with our notation is simply $\alpha = 1/b(\sigma,\nu)$ and $\beta = \nu$.

We differentiate f to get

$$\frac{f(t)}{f'(t)} = -\frac{\alpha^{\beta}}{\beta} t^{1-\beta}.$$

Taking the limit for $t \to \infty$ we can distinguish three cases:

(a) $\beta > 1$: This case includes the Gaussian density ($\beta = 2$); the slope bound is $D'(0) \leq \lambda_0 = 0$. This is trivially true for any distortion rate function and hence not very useful. Since λ_0 is the slope of bound (2.59) at $R = 0$, this result actually tells us that for $\beta > 1$ that bound is really bad. Because $\lambda_0 = 0$ implies that the bound is not even convex and could be improved by timesharing (multiplexing) between $(D, R) = (\sigma^2, 0)$ and some point with $R > 0$.

(b) $\beta = 1$ (Laplacian density): $D'(0) \leq \lambda_0 = -2\alpha^2 = -\sigma^2$.

(c) $\beta < 1$: We have $D'(0) \leq \lambda_0 = -\infty$, which means that $D(R)$ decays very rapidly at low rates.

Thus Theorem 2.6 suffices to establish that for generalized Gaussians with $\beta < 1$, i.e. those which are more peaked than a Laplacian, the D axis is tangent to $D(R)$ at

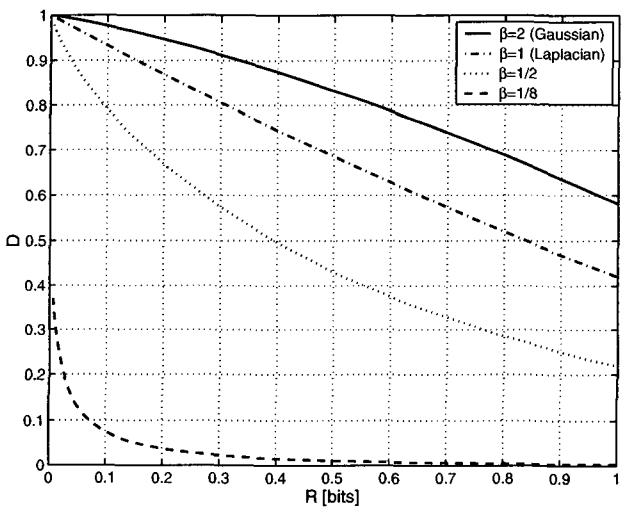

Figure 2.5: Weakened upper bound Eq. (2.59) for generalized Gaussian $D(R)$.

$(R, D) = (0, \sigma^2)$. This discontinuous behavior of the slope bound as a function of β is surprising at first, since the family of generalized Gaussian densities is continuous in the shape parameter β. However, if we consider a single density $f_{\alpha,\beta}(x)$ in this family, we see that the tail decay will be super-exponential for $\beta > 1$ and sub-exponential for $\beta < 1$. This is the cause for the abrupt slope change. Figure 2.5 plots the bound (2.59) for several values of β; the corresponding densities are shown in Figure 2.4.

A good compromise We have seen two approximations to $R^*(t)$ that lead to upper bounds with different appeals: precision for (2.57), simplicity for (2.59). Since allegedly the Swiss always strive for a compromise, we will try to find one by approximating the argument of the second logarithm in (2.56). Recall that $h'(\mu) = \ln(1-\mu) - \ln(\mu)$. For large t (thus small μ) the term $-\ln \mu$ dominates $\ln(1-\mu) \approx -\mu$.

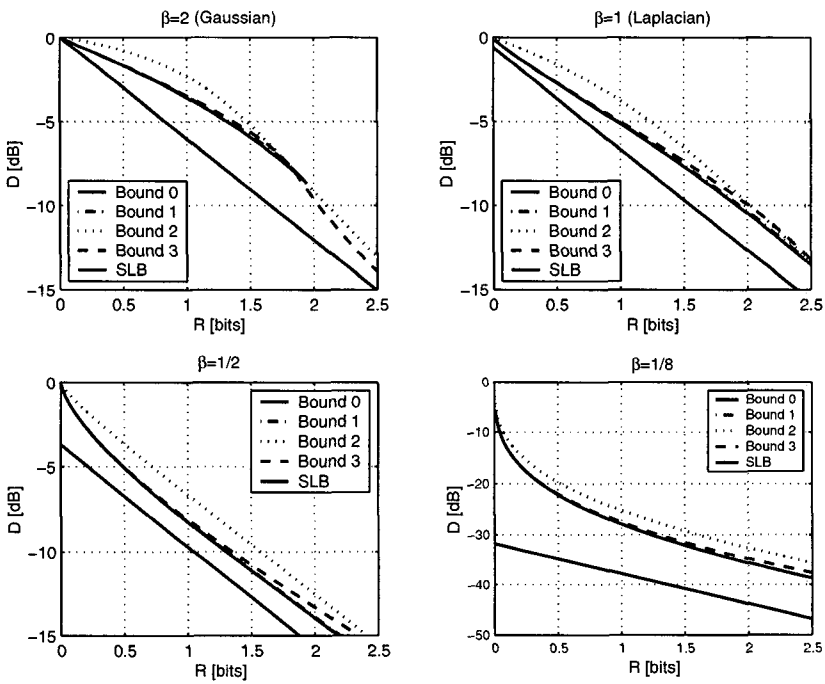

Figure 2.6: Comparison of upper bounds for generalized Gaussian $D(R)$. Legend: "Bound 0" is the magnitude bound (2.18), "Bound 1" is Eq. (2.57), "Bound 2" is Eq. (2.59) and "Bound 3" is Eq. (2.64); "SLB" is the Shannon lower bound.

In fact, for large enough t, $\gamma(t)$ is close to one, so that $-\ln \mu$ dominates all terms. Mutatis mutandis $-\ln \gamma$ would become important for small t. However we are more interested in tight bounds at low rates, that is for large t, and therefore our compromise rate is

$$R^*(t) \approx R_3(t) = h(\mu(t)) + \frac{\mu(t)}{2}\left[-\ln \gamma(t) + \ln\left(-2 \ln \mu(t)\right)\right] \tag{2.63}$$

and the corresponding upper bound is

$$D(R_3(t)) \leq \frac{A(t)\gamma(t)}{-2 \ln \mu(t)} + \sigma^2 - A(t) = \frac{\mu(t)\, t^2}{-2 \ln \mu(t)} + \sigma^2 - A(t). \tag{2.64}$$

This is tighter than (2.59) for all $\mu < e^{-1/2}$ (but also the rate has been slightly increased).

All the three weakened bounds and the magnitude upper bound are shown in Figure 2.6 for four different densities of the generalized Gaussian family. As the shape parameter gets smaller, the density is more peaked (see Fig. 2.4) and the low-rate decay of the bounds becomes steeper. The magnitude bound (2.18) and Eq. (2.57) are almost indistinguishable, whereas the weakest bound, Eq. (2.59), honors its appellation. Finally, Eq. (2.64) indeed strikes a good compromise, since as promised it is quite tight at low rates.

2.6 The Flat-Peaked Random Variable

In this section we are going to analyze a simple example of a random variable that is peaked at its mean and has heavy tails, as shown in Figure 2.7. A proper pdf $f(x)$ has to satisfy

$$\int f(x)\, dx = 2(ad + bc - ac) = 1. \tag{2.65}$$

Further, we will normalize the variance to one, or

$$\int x^2 f(x)\, dx = 2\left(d\frac{a^3}{3} + c\frac{b^3 - a^3}{3}\right) = 1. \tag{2.66}$$

With these two equations we can easily compute c and d to be

$$c = \frac{3 - a^2}{2b(b^2 - a^2)}, \tag{2.67}$$

$$d = \frac{a^2 + ab + b^2 - 3}{2ab(a + b)}. \tag{2.68}$$

From the constraints $b > a > 0$ and $d > c > 0$ we get the ranges of a and b for which the box peak pdf is well defined: $0 < a < \sqrt{3}$ and $b > \sqrt{3}$.

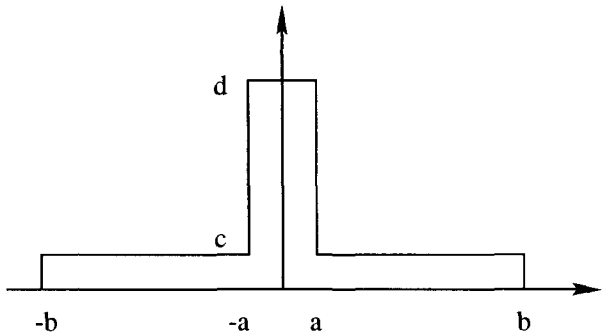

Figure 2.7: The pdf of a flat-peaked random variable

The goal of this example is to compare the upper bound to different scalar quantization approaches, which are easy to analyze for the piecewise uniform density at hand. The pdf is determined by any two parameters: we choose the width of the peak, $2a$, and its probability mass $p_0 = 2ad$. This situation can also be characterized by saying that $2a$ is the smallest volume of any (one-dimensional) set of probability p_0.

At low rates, we consider symmetric three, resp. five level scalar quantizers. The width of the center bin, which is equal to twice the threshold, is varying from $2t = 2a \ldots 2b$. For the outside bins, a Lloyd-Max quantizer for the *thresholded* random variable is used. In general, the reconstruction points are the centroids of the outside bins. However, in the particular case $t \geq a$ that we consider, the Lloyd-Max quantizer reduces to a uniform quantizer, where the reconstruction points are identical to the bin centers. This is a special case of Uniform Threshold Quantization (UTQ) analyzed in [15]: one center bin of width $2t$, and $2N$ surrounding bins of width $(b-a)/N$ each.

The first two curves in Figure 2.8 show the achievable performance: the local minima correspond to the three, resp. five level Lloyd-Max quantizers. We recall that the quantizer indices are entropy coded to achieve rates below one bit. The endpoints on the right coincide with a quantizer with zero bin $]-a, a[$ plus two, resp. four uniform outside bins. The third curve is for a scheme that optimally allocates rate among uniform quantizers for each of the three regions $[-b, -a]$, $]-a, a[$ and $[a, b]$. This again corresponds to UTQ, but with the added freedom of allowing a non-integer number of bins $N_i = 2_i^R$. Note that the minimum rate will be $H(1-p_1, p_1/2, p_1/2 = H(p_1) + p_1$ bits, i.e. when each region corresponds to one quantization bin ($p_1 = 1 - p_0$). As is obvious from the figure, this scheme is far from optimal at low rates. The other curves in Figure 2.8 are the magnitude upper bound, the empirical $D(R)$ computed with Blahut's algorithm and the SLB. The knee in the upper bound is a good indicator for the beginning of the high-rate region, where the bound becomes loose.

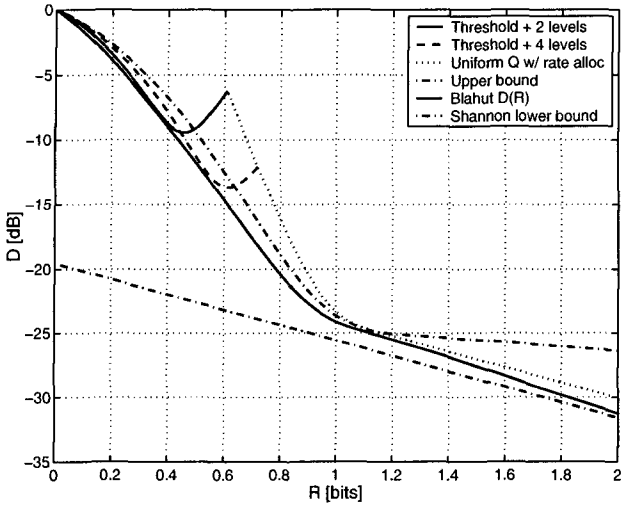

Figure 2.8: Distortion rate characterization of flat-peaked source. Parameters $a = 0.1, p_0 = 1 - 0.11$.

The differential entropy needed to compute the SLB can be expressed as

$$h(X) = -\int f(x) \log f(x)\, dx = -2ad\log(d) - 2(b-a)c\log(c) \quad (2.69)$$

$$= H(p_1) + p_0 \log(2a) + p_1 \log(p_1^{-1/2}\sqrt{12 + (p_1 - 4)a^2} - 3a), \quad (2.70)$$

where we used $b = (1/2)(-a + \sqrt{12p_1^{-1} + (1 - 4p_1^{-1})a^2})$. Equation (2.70) makes evident that for $p_0 \to 1$ (therefore $p_1 \to 0$) the term $\log(2a)$ dominates all others, since $\lim_{p\to 0+} p \ln p = 0$. This underlines the importance of the width (the volume) of the most probable set, which essentially determines the entropy.

2.6.1 Infinite Tails

Now we will fix the value of a and let $b \to \infty$, respectively $p_1 \to 0+$. This corresponds to a random variable that concentrates essentially all probability to the finite interval $]-a, a[$, but has infinite tails such that the variance constraint (2.66) is satisfied. First, the the rate distortion behavior will be characterized by the Shannon lower bound and the magnitude upper bound. Then the performance of a practical uniform quantization scheme is analyzed.

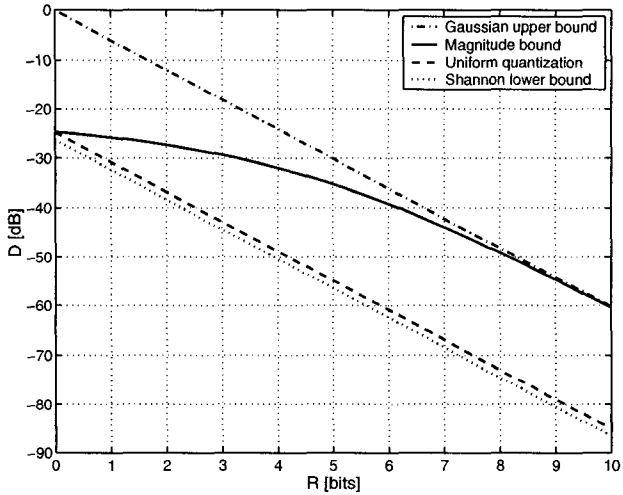

Figure 2.9: Distortion rate characterization of flat-peaked source for $b \to \infty$.

Shannon lower bound As expected from (2.70), we have

$$\lim_{b\to\infty} h(X) = \lim_{p_1\to 0+} h(X) = \log(2a).$$

The more confined the peak (smaller a), the lower the entropy and thus the SLB. If we take a as an (inverse) measure of peakedness, this simply means that highly peaked random variables have lower asymptotic (high-rate) distortion than less peaked ones, i.e. they are easier to compress.

Magnitude upper bound First we treat the case for thresholds $t > a$. Define $t = \tau b$, so that for $b \to \infty$ the bound can be computed for all $0 < \tau < 1$. The key quantities needed to compute the bound are

$$\mu(\tau) = \frac{3 - a^2}{b^2 - a^2}(1 - \tau), \tag{2.71}$$

$$A(\tau) = \frac{3 - a^2}{3 - 3a^2/b^2}(1 - \tau^3), \tag{2.72}$$

Since $\lim_{b\to\infty} \mu(\tau) = 0$ regardless of τ, we have that for any $R > 0$ the exponential $\exp(-2[R - h(\mu)]/\mu)$ in the bound (2.18) goes to zero, implying $D(R > 0) \leq \sigma^2 - A(\tau)$. Furthermore, $\lim_{b\to\infty} A(\tau) = (1 - a^2/3)(1 - \tau^3)$, which means the bound is tightest for $\tau \to 0$. Therefore we can bound the distortion of the flat-peaked

source with

$$D(R) < \frac{a^2}{3} + \delta \qquad \forall R > 0 \tag{2.73}$$

for an arbitrarily small positive δ. Thus $D(R)$ drops abruptly from 1 to $a^2/3$ (or less) when the rate is increased from 0.

Now let's look at $t \leq a$, where after taking the limit $b \to \infty$ we have

$$\mu(t) = 1 - t/a, \tag{2.74}$$

$$A(t) = 1 - t^3/3a. \tag{2.75}$$

These can be directly plugged into one of the upper bounds. If we use (2.58) and (2.59) and set $t = a$, we get $R(a) = 0$ and $D(R(a)) \leq a^2/3$ as expected. For Figure 2.9 the tighter bound (2.64) was used. By taking the limit $b \to \infty$ we have actually reduced our flat-peaked pdf to a random variable that with probability one is uniformly distributed over $[-a, a]$ — not very interesting.

2.6.2 Uniform Spike

Finally we analyze the case $a \to 0+$, but keeping the probability mass p_0 at zero. This corresponds to having a Dirac distribution $\delta(x)$ of weight p_0 plus an uniform density on $[-b, b]$ with probability $1 - p_0$. This setup actually anticipates the spike process that will be analyzed in more detail in the next chapter, only that there the continuous density will be Gaussian, not uniform.

The moment constraints (2.65, 2.66) require

$$b = \frac{\sqrt{3}}{\sqrt{1-p_0}} \text{ and } c = \frac{(1-p_0)^{3/2}}{2\sqrt{3}}.$$

With these equations we obtain the moment functions $\mu(t) = (1 - p_0)(1 - t/b)$ and $A(t) = 1 - t^3/b^3$, from which we can plot the magnitude bound. The distortion for a uniform quantizer is straightforward to compute if we observe that it operates only a proportion $1 - p_0$ of the time. Its operational drf is

$$\delta(R) = \exp(-2\tfrac{R-h(p_0)}{1-p_0}), \qquad R \geq h(p_0). \tag{2.76}$$

Figure 2.10 shows that we can improve on (2.76) by linear multiplexing between the $(R, D) = (0, 1)$ and the point on $\delta(R)$ that has the tangent through $(0, 1)$. The other remarkable point is that the bound is actually better than a uniform quantizer, even when multiplexing with $R = 0$.

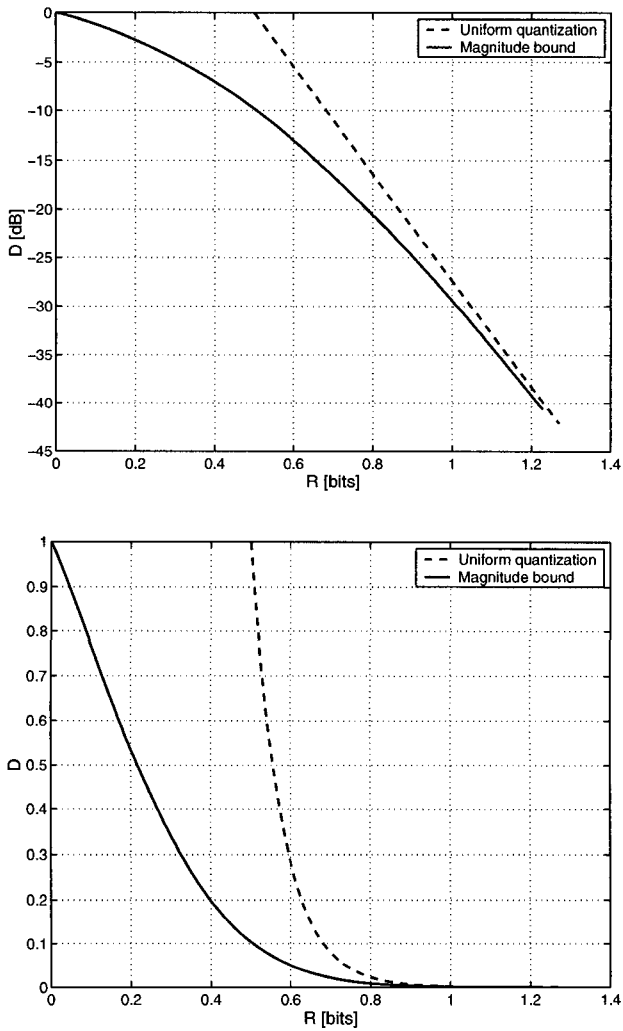

Figure 2.10: Distortion-rate upper bound and uniform quantizer performance for uniform spike source. Linear scale at bottom.

Chapter 3

Low-Rate Transform Coding

In this chapter we are going to address the question that first motivated our research: what makes transform coding systems do so well at low rates? We will take an ad hoc approach with a slant towards wavelet transforms; in particular, all examples will be for wavelet image compression. As a consequence, we will only be able to give partial answers to the opening question. Nevertheless, the bounds from this and the previous chapter can be used to distinguish between transforms with coefficient statistics that are "good" or "bad" for compression. In that sense these bounds are engineering tools that quantify one's intuition about low-rate transform coding.

The outline of this chapter is as follows: first, we will argue that "good" transforms output coefficient vectors that are sparse, i.e. most components are equal to zero or close to it. This observation leads to the definition of the *spike process*, which is an extremely simple model for sparse coefficients. If we are just concerned with the positions of the nonzero components, the Hamming distortion measure is appropriate. Section 3.2 is an exception to the general flavor of this thesis, since it presents *exact* analytic solutions for this special Hamming case. In Section 3.3 we get back to upper bounds on mean squared error when we look at Gaussian-valued spike processes. As it turns out, the spike model is not a good match for the behavior that we actually see in practical systems. Therefore in Section 3.4 we look at Gaussian mixture coefficient models, which will give more satisfying results. They also expose the weakness of the magnitude bound at higher rates. We take that as an incentive to derive a high-rate upper bound that complements the low-rate bound, but at the price of giving up on oligoquantization. In exchange the high-rate bound also allows us to define a coding gain similar to the one known in linear transform coding. This opens a new perspective on those transform coding systems that "mix" all coefficients together and quantize them with a single quantizer, which is matched to the mixture density.

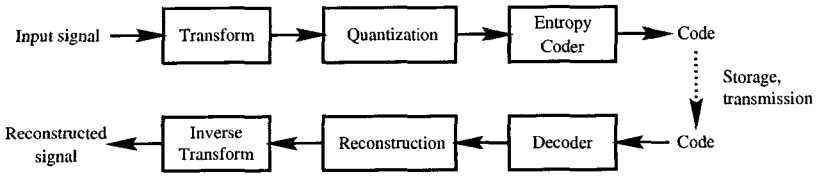

Figure 3.1: Block diagram of a transform coding system

3.1 Sparse Transform Coefficients

The basic block diagram of a transform coder is shown in Figure 3.1. In most cases the signal is split into blocks of constant length before it is transformed and quantized. The quantizer is usually a direct product of scalar quantizers: in fact, the main "raison d'être" of transform coding is to avoid the complexity of vector quantization. If the transform is orthogonal, Parseval's equality holds and the distortion in the signal domain will be equal to the quantizer distortion in the transform domain. This is also approximately true for some often used biorthogonal wavelet transforms, since they have almost tight frame bounds. Thanks to this relation, the design of the transform and the quantizer can be separated.

The goal of the transform is to decorrelate the signal, or more generally, to pack the signal energy in as few coefficients as possible. If the input signal is a stationary Gaussian process, then the Karhunen-Loève transform (KLT) is known to be the best such transform. The best high-rate product quantizer can then be found with standard bit allocation methods. Even a low-rate analysis of this Gaussian/KLT case is possible [30].

The problem with these approaches is that most natural signals are highly non-Gaussian, so the KLT analysis does not apply. Nevertheless, transform coding works! We sketch an explanation of this for the wavelet transform. Figure 3.2 shows a dyadic one-dimensional wavelet transform realized with an iterated filter bank. H_1 is a highpass filter that outputs the *details*, whereas the lowpass H_0 outputs a coarse *approximation* of the input signal. Each time we iterate over the lowpass output, we are analyzing the signal at a larger *scale*.

A key property of a wavelet is the number of vanishing moments. In a nutshell, it says up to what degree a polynomial will be canceled in the highpass (and thus be represented entirely by the lowpass aka scaling coefficients). Figure 3.3 shows the wavelet decomposition of a piecewise polynomial signal. If the wavelet has enough vanishing moments, only a fixed number of nonzero detail coefficients will appear at each scale around the signal singularities. The polynomial part will be entirely represented by the lowpass approximation. This behavior extends to piecewise regular signals and is in fact a good first approximation for wavelet transforms of arbitrary signals that are mainly composed of smooth pieces [14]. These are good models not

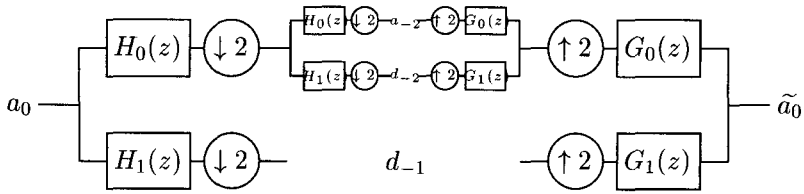

Figure 3.2: Wavelet transform with an iterated filter bank.

only for images[1], but also many other non-stationary, non-Gaussian natural signals. Here we have to remark that our "derivation" holds in one dimension, whereas the situation is less straightforward for the separable two-dimensional transforms used in image compression. However, the observed coefficient behavior still supports the hypothesis that the wavelet coefficients describe singularities, while the scaling coefficients approximate the smooth signal component. There is simply a certain redundancy (an increased number of nonzero coefficients), since a separable transform cannot "recognize" the proper singularities (i.e. edges) in two dimensions.

To sum up, the wavelet transforms used in compression yield vectors of sparse large amplitude coefficients. They live up to the goal of packing as much energy as possible in as few coefficients as necessary. In fact even transforms of the Fourier family show a similar behavior if they operate on small blocks, like in the JPEG image coding standard. Another example are the coefficients in LPC speech coding when the sound is voiced.

Of course these sparse transform vectors are an ideal match for oligoquantization, since the thresholding will retain precisely the large amplitude coefficients that contain most of the signal energy. Note that now we consider a single (scalar) quantizer for *all* coefficients, and no longer the product quantizers with bit allocation that are used with the KLT. This makes sense because we cannot predict *a priori* the positions of large coefficients[2], so that precomputed bit allocation is impossible.

In this chapter we will analyze two models for sparse transform coefficients. First, we propose a strongly simplified model of a transform vector that is zero everywhere except in a few positions, where large "spikes" stick out (see the detail scales in Figure 3.3). This spike model will be studied in the next two sections. However, due to its discontinuous nature, it does not match the behavior of actual transform coding systems. Therefore we then look at Gaussian mixtures, which prove to be a much better model.

[1] Except for high frequency textures.

[2] Of course causal prediction across wavelet scales is possible. However the position of a singularity in the signal cannot be predicted.

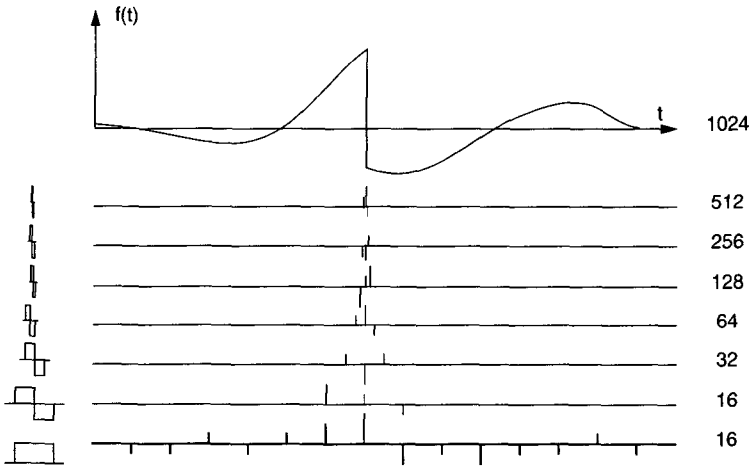

Figure 3.3: Sketch of the dyadic wavelet transform of a piecewise polynomial signal: six detail (wavelet) scales plus lowpass approximation. Left: example wavelets and scaling function; right: example for the number of samples (top), resp. coefficients.

Literature Transform coding is treated in almost every textbook that covers lossy data compression, see e.g. [26, 28, 43]. Wavelet-oriented signal processing textbooks that contain material on transform coding are e.g. [47, 32]. An analysis of the low-rate behavior of image transform coders, based on measured statistics, appeared in [33].

3.2 Spike Position Encoding

Before we generalize to random spikes in the next section, we want to analyze a restricted subproblem that is *deterministic* in the sense that the number of spikes is known a priori to both encoder and decoder. For the Hamming distortion measure it is possible to obtain analytic expressions for the rate distortion function.

3.2.1 Definitions and Single Spike Case

In the following definition the spike location will be restricted to a finite set. This makes sense, not only because in practice one always deals with a finite number of transform coefficients, but also because a continuous distribution of spike locations would require infinite rate for zero distortion.

Definition 3.1 *A **spike process** is a memoryless random process which outputs or-*

dered pairs (U, V) of random variables, where the position U *has a finite alphabet $\mathcal{U}$ with $|\mathcal{U}| = N$, while the* value V *is drawn from an alphabet $\mathcal{V}$ with no particular restrictions.*

We will assume $\mathcal{U} = \{1, 2, \ldots, N\}$, since any other alphabet of size N can always be represented by that set. Further we suppose that $\mathcal{V}$ is a ring. Then a spike process sample (u, v) can be mapped to a $\mathcal{V}^N$-vector simply by $\phi(u, v) = v\mathbf{e}_u$, where $\mathbf{e}_i$ is the i-th standard basis vector.

Definition 3.2 *Let $\mathcal{X}$ be the source alphabet, and $\widehat{\mathcal{X}}$ the reconstruction alphabet. A* ***single-letter distortion measure*** *(or fidelity criterion) is a non-negative real-valued function $\rho(x, \hat{x}) : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, \infty[$. In the spike case we have $x = (u, v)$, $\mathcal{X} = \mathcal{U} \times \mathcal{V}$, and analogously $\hat{x}$, $\widehat{\mathcal{X}}$.*

Probably the simplest distortion measure that can be applied to spike processes is the Hamming distance between the position vectors, i.e. the basis vectors $\mathbf{e}_i$. It can be shown (see Theorem 3.1 below) that just one additional reconstruction letter is needed to achieve the rate distortion bound, and it will map to the all-zero vector $\mathbf{0}$. If we define $\widehat{\mathcal{U}} = \{0\} \cup \mathcal{U}$ and $\mathbf{e}_0 = \mathbf{0}$, then everything fits nicely. Using $\hat{u} = 0$ corresponds to not coding the position. We get the following distortion measure:

$$\begin{aligned} \rho(u, v; \hat{u}, \hat{v}) &= w_H(\mathbf{e}_u - \mathbf{e}_{\hat{u}}) \\ &= \delta(\hat{u}) + 2[1 - \delta(\hat{u})][1 - \delta(u - \hat{u})] \end{aligned} \tag{3.1}$$

Thus "giving the right answer" has zero distortion, a wrong answer two, and not answering costs one distortion unit. We proceed with the definition of the (information) rate distortion function for discrete memoryless sources (DMS), closely following the notation in Berger's book [1]. Note that this coincides with the information-theoretic definition ("infimum of achievable rates").

Definition 3.3 (Rate distortion function of a DMS) *Let $X \sim P$ be a discrete memoryless random variable, $\rho(x, \hat{x})$ a single-letter distortion measure, $Q_{\widehat{X}|X}(k|j)$ a conditional distribution (defining a random codebook), and $P_{X,\widehat{X}}(j, k) = P(j)Q(k|j)$ the corresponding joint distribution. The* average distortion *associated with $Q(k|j)$ is*

$$d(Q) = \sum_{j,k} P(j)Q(k|j)\rho(j, k). \tag{3.2}$$

If a conditional probability assignment satisfies $d(Q) \leq D$ it is called D-admissible. *The set of all D-admissible Q is $Q_D = \{Q(k|j) : d(Q) \leq D\}$. The average mutual information ("description rate") induced by Q is*

$$I(Q) = \sum_{j,k} P(j)Q(k|j) \log \frac{Q(k|j)}{Q(k)}, \tag{3.3}$$

where $Q(k) = \sum_j P(j)Q(k|j)$. *The rate distortion function* $R(D)$ *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 [1], [10, Section 13.7]. We start with the functional

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

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 can be computed as

$$Q(k|j) = \frac{Q(k)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}}. \tag{3.4}$$

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

$$\sum_j \frac{P(j)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}} = 1 \qquad \text{if } Q(k) > 0, \tag{3.5}$$

$$\sum_j \frac{P(j)e^{-\lambda\rho(j,k)}}{\sum_{k'} Q(k')e^{-\lambda\rho(j,k')}} \leq 1 \qquad \text{if } Q(k) = 0. \tag{3.6}$$

Inequality (3.6) stems from the Kuhn-Tucker conditions (for a detailed derivation of the above see Section 13.7 in [10]). The solution of the problem is further simplified by the following theorem (2.6.1 in [1]):

Theorem 3.1 (Berger) *No more than* N *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,* $\widehat{N} = N + 1$ *reproducing letters are needed for a point that lies on a straight-line segment.*

As anticipated above, we need just one additional reconstruction letter to achieve $R(D)$. To see that it can only be the all-zero vector, consider the source alphabet $\widehat{\mathcal{U}} \cong \{e_{\widehat{\mathcal{U}}}\}$, 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, i.e. exactly one.

Proposition 3.2 *The rate distortion function for a single spike in* $N \geq 2$ *equiprobable positions with the Hamming distortion criterion (3.1) is*

$$R(D) = \begin{cases} (1-D)\log(N-1) & \text{if } \frac{2}{N} < D \leq 1, \\ \log N - \frac{D}{2}\log(N-1) - h\left(\frac{D}{2}\right) & \text{if } 0 \leq D \leq \frac{2}{N}. \end{cases} \tag{3.7}$$

Proof: The symmetry of the input distribution, $P(j) = 1/N$ $(j = 1, \ldots, N)$, suggests the following marginal distribution:

$$Q = (q_0, q_1 = q_2 = \ldots = q_N = \frac{1 - q_0}{N}). \tag{3.8}$$

Let us first assume that $q_k > 0$ holds for all k. Then the $N + 1$ conditions (3.5) have to be met. We make the substitution $\beta = e^{-\lambda}$ and insert our $Q(k)$ into the equation, first for $k \neq 0$:

$$\begin{aligned} \frac{\beta^0}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} + \frac{(N-1)\beta^2}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} &= \frac{1}{P(j)} = N \\ &\vdots \\ q_0((N-1)\beta^2 - N\beta + 1) &= 0. \end{aligned} \tag{3.9}$$

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

$$\begin{aligned} \frac{N\beta^1}{q_0\beta^1 + \frac{1-q_0}{N}(\beta^0 + (N-1)\beta^2)} &= \frac{1}{P(j)} = N \\ &\vdots \\ (1 - q_0)((N-1)\beta^2 - N\beta + 1) &= 0. \end{aligned} \tag{3.10}$$

The solution $\beta = 1$ corresponds to the point $(0, D_{max})$ (with $D_{max} = 1$) in the (R, D) plane, which is achieved by setting $q_0 = 1$. Therefore the interesting solution is $\beta = 1/(N-1)$, which when inserted into (3.4) yields

$$Q(k|j) = q_k(N-1)^{1-\rho(j,k)}. \tag{3.11}$$

Putting (3.11) into (3.2) we get the average distortion $d(Q) = 1 - \frac{N-2}{N}(1 - q_0)$ and from (3.3) 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

$$D(R) = 1 - \frac{R}{\log(N-1)} \qquad \text{for } R < \frac{N-2}{N}\log(N-1). \tag{3.12}$$

This proves the first part of equation (3.7). When R reaches its upper bound in (3.12), D reaches $2/N$ and we have $q_0 = 0$. At that point, equation (3.9) will be satisfied for all β. According to condition (3.6), equation (3.10) now becomes an inequality:

$$(N-1)\beta^2 - N\beta + 1 \geq 0. \tag{3.13}$$

This is satisfied by $\beta \geq 1$ or $\beta \leq \frac{1}{N-1}$, which is equivalent to $\lambda \geq \ln(N-1)$. The first solution ($\beta \geq 1$) can be discarded, since it would result in $D(R)$ being larger than 1 and discontinuous. The conditional distribution parameterized by β is

$$Q(k|j) = \begin{cases} 0, & k = 0 \\ \frac{\beta^{\rho(j,k)}}{1+(N-1)\beta^2}, & k \neq 0 \end{cases} \tag{3.14}$$

As before we put this into (3.2) to get $d(Q) = \frac{2(N-1)\beta^2}{1+(N-1)\beta^2}$ and into (3.3) yielding

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

where $h(p) = -p\log p-(1-p)\log(1-p)$ is the binary entropy function. Eliminating β from the last two equations yields the second part of equation (3.7). □

Figure 3.4 shows a set of typical $R(D)$ functions. As N grows large, the linear segment dominates the rate distortion characteristics. Further we observe that in the special case $N = 2$ the solution degrades to the $D(R)$ function of a binary symmetric source (with doubled distortion).

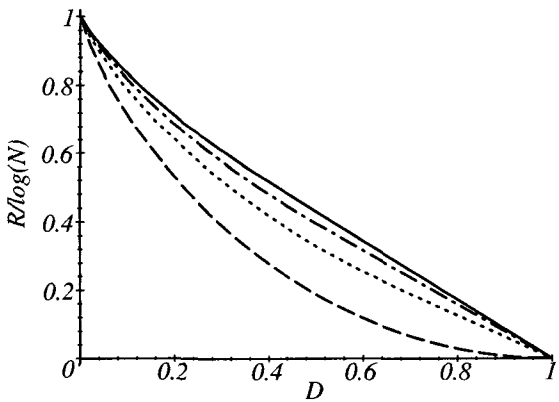

Figure 3.4: $R(D)$ for single spike with Hamming distortion, $N = 2$ (bottom) up to $N = 5$ (top curve). The rate has been normalized to $1/\log N$. For $N \to \infty$, $R(D)$ becomes a straight line, see (3.7).

3.2.2 Generalizing to Multiple Spikes

The typical output of an image transform is of course not a single spike in one out of N positions, but rather a group of them. Therefore a way to model a transformed picture is to specify K out of N coefficient positions, e.g. those above a given threshold. In the Hamming case this is equivalent to a source emitting one of the $\binom{N}{K}$ binary vectors of length N and Hamming weight K. We will again assume that all source letters are equally probable, and that N and K are given. We look only at the case where the number of 1's is $K \leq N/2$, since the other case ($N/2 \leq K \leq N$) is complementary.

The analysis is simplified by the fact that the set of source vectors of weight K forms a group code under permutation. Under the action of the symmetric group S_N,

any vector of the set will again yield the whole set. Thus the code is geometrically uniform, i.e. the distance (distortion) profile looks the same from any vector of the set. By decomposing permutations into transpositions, one establishes that the distances will always be integer multiples of two. Assuming that $K \leq N/2$, there are exactly

$$w_d = \binom{K}{d}\binom{N-K}{d}, \quad d = 0, \ldots, K \tag{3.15}$$

vectors at Hamming distance $2d$ from a given vector. The following identity will also be very helpful in our development:

$$\sum_{d=0}^{K} w_d = \sum_{d=0}^{K} \binom{K}{d}\binom{N-K}{d} = \binom{N}{K}. \tag{3.16}$$

As in the single spike case, the reconstruction alphabet consists of the source alphabet plus the zero vector, to which we assign the probability q_0 as before. To compute the slope of the linear part of the rate distortion curve we have to solve the equation (compare with (3.9, 3.10))

$$\sum_{d=0}^{K} w_d \beta^{2d} - \binom{N}{K}\beta^K = 0. \tag{3.17}$$

The solution $\beta = 1$ corresponds again to maximum distortion, $D = K$. We will now assume that somehow we found the interesting root β_0 with $0 < \beta_0 < 1$ (for $K = 2$ it is $\beta_0 = \binom{N-2}{2}^{-1/2}$, for larger K it can be computed numerically). Then the linear part of the rate distortion function will be

$$R(D) = (D - K)\log\beta_0, \quad D(\beta_0) < D < K \tag{3.18}$$

where the bounds on D guarantee $0 < q_0 < 1$ ($D(\beta_0)$ is defined below in (3.20a)). For $q_0 = 0$, any $\beta \leq \beta_0$ will satisfy the Kuhn-Tucker conditions. We define a pseudo-distribution

$$b_d = \frac{w_d \beta^{2d}}{\sum_{d'=0}^{K} w_{d'} \beta^{2d'}}, \quad d = 0, \ldots, K. \tag{3.19}$$

After some calculations, we get a parametric expression for the rate-distortion curve:

$$D(\beta) = \sum_{d=1}^{K} b_d 2d \tag{3.20a}$$

$$R(\beta) = \log\binom{N}{K} + \sum_{d=0}^{K} b_d \log b_d - \sum_{d=0}^{K} b_d \log w_d \tag{3.20b}$$

for $0 < \beta < \beta_0$. The middle term in the expression for R is the negative entropy of our pseudo-distribution b_d (compare with (3.7)). Figure 3.5 shows that for sparse spikes

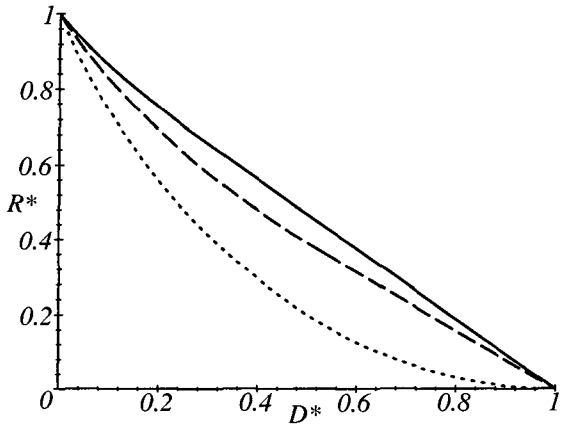

Figure 3.5: $R(D)$ for multiple spikes with Hamming distortion, $K = 4, 8, 16$ spikes (top to bottom curve) in $N = 32$ positions. Rate, distortion normalized to $R^* = R/\log\binom{N}{K}$ and $D^* = D/K$.

(small K/N) the linear segment again dominates the rate distortion behavior. The consequence of this "almost linear" $R(D)$ behavior for sparse spikes is the following: to build a close to optimal encoder for intermediate rates $0 < R < \log\binom{N}{K}$, we can simply multiplex between a rate 0 code (no spikes coded) and one with rate $\log\binom{N}{K}$ (all K spikes coded exactly). Put otherwise, if we have a bit budget to be spent in coding a sparse binary vector, we can simply go ahead and code the exact positions of the ones (the spikes) until we run out of bits.

These results can also be used to derive the asymptotic operational rate distortion function of a simple two pass universal lossy source coder as follows: first the number of ones (K) in a block of length N is determined and sent to the decoder using at most $\log_2 N$ bits. Then a code for a weight K vector is used. For $N \to \infty$, we approach the above rate distortion functions with a redundancy of $\frac{\log_2 N}{N}$ bits per sample. (In view of the results for universal lossless source coding, we expect that this redundancy could be halved.)

3.3 Random Spike Processes

The previous section was a short excursion to Hamming distortion and precise results: now we are back to mean squared error and upper bounds. Instead of considering only the spike positions, now we also look at their values. Moreover, we abandon the deterministic setting "K spikes in N positions" in favor of a memoryless, scalar model. Figure 3.6 shows the basic setting: a spike process is simply the product of

a binary-$\{0,1\}$ source with a memoryless real-valued source. Here we consider only Gaussian values, because they serve as the usual worst-case benchmark. The binary source simply "switches the value source on or off". To start, we consider a BMS, so that the resulting spike process is also memoryless.

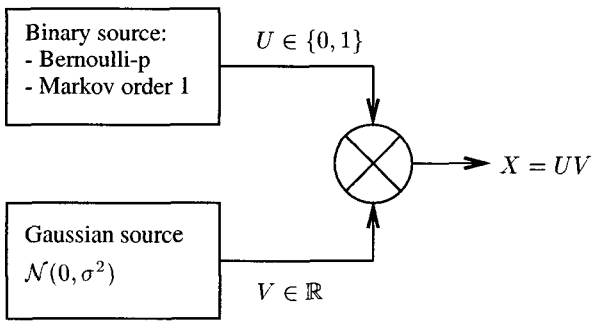

Figure 3.6: Spike process

Definition 3.4 (Bernoulli-Gaussian (BG) spike process) *An i.i.d. Bernoulli-Gaussian spike source (process) emits a memoryless random variable X that is the product of a binary random variable with a zero mean Gaussian random variable. The binary (Bernoulli) random variable has $\Pr\{X = 1\} = p$ and $\Pr\{X = 0\} = 1 - p$, and the Gaussian has variance σ^2. Using the $\delta(\cdot)$ distribution, the pdf of the BG spike can be written as*

$$f(x) = (1-p)\delta(x) + p\frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/2\sigma^2}. \tag{3.21}$$

This pdf can also be seen as a mixture of two zero mean Gaussian random variables, with one of them having zero variance (a special case of the model that will be studied in the next section). The magnitude bound (2.18) can be easily evaluated if one replaces T by $T + \epsilon$ in the lower integration boundary of (2.2), with an arbitrarily small number $\epsilon > 0$. By doing this we exclude the Dirac $(1-p)\delta(x)$ from the integral, and hence we have $\mu(T) \leq p$ for all $T \geq 0$. This is obviously correct, since we never have to code the value of a spike with zero amplitude.

Figure 3.7 shows the bound and the empirical $D(R)$ for $p = 0.11$. The asymptote shown is actually the trivial upper bound $B(0,R)$, i.e. when all spikes are coded (thus at least $R = h(0.11) = 0.5$ bits are required before the distortion starts decreasing). The figure illustrates the change in $D(R)$ behavior between low and high rates that is typical of spike processes, regardless whether the continuous part of the pdf is a Gaussian or some other density (see also the uniform spike in Section 2.6.2).

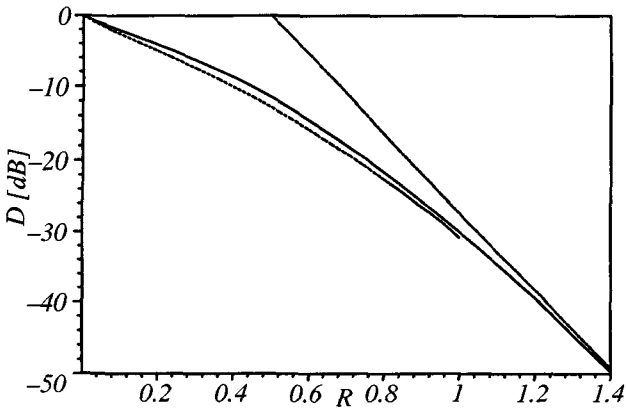

Figure 3.7: Bernoulli-Gaussian spike with $p = 0.11$: empirical $D(R)$, upper bound (2.18) and trivial upper bound $B(0, R)$ (bottom to top curve). Normalized to unit variance.

Mixed random variables When evaluating the magnitude bound for the BG spike, we excluded the Dirac $\delta(x)$ from the integrals. The deeper reason is quite simple: most results in "standard" rate distortion theory hold only for random variables with (absolutely) continuous densities, some hold if there is at most a countable number of steps in the pdf. That the BG spike is a strange random variable is also evident from the plotted bounds, since the asymptotic distortion decay is much steeper than -6 dB per bit. In fact, the spikes belong to the family of *mixed* random variables, that have both a discrete and a continuous part. Their entropy cannot be computed with the usual integral, but only via mutual information conditioned on the discrete part [37, Ch. 2]. With this trick, Rosenthal and Binia were able to derive the asymptotic rate distortion behavior of mixed random variables [39]. Their results coincide with our asymptotic upper bound $B(0, R)$ if the continuous part is Gaussian, otherwise their result is obviously tighter.

3.3.1 First Order Markov Spikes

A natural extension of the memoryless spike model is to consider bursts of spikes. To model such a bursty behavior we can simply replace the Bernoulli process by a first order binary Markov process, with states S_0 (no spike) and S_1 (spike). Thus we get a spike process with correlated positions, but uncorrelated values. The Markov process is specified by two transition probabilities, $\alpha = \Pr\{S_0 \to S_1\}$ and $\beta = \Pr\{S_1 \to S_0\}$.

To compute a bound like (2.18) we first threshold the Markov-Gaussian spike process. It is easy to show that the resulting process is the product of a new first order

Markov process and a *thresholded* i.i.d. Gaussian process. The parameters of the new ("thresholded") Markov process can be computed from α, β and the threshold t. Then we compute the entropy rate of the thresholded process, i.e. the spike position indicator (2.3), and use it to replace $h(\mu(T))$ in equation (2.16).

Figure 3.8 compares upper bounds for memoryless and Markov spikes with the same marginal spike probability, $p = \alpha/(\alpha + \beta)$. At low rates the two bounds are quite close, while at higher rates their horizontal spacing approaches the difference between zero-th and first order entropy (entropy rate). The low-rate behavior can be explained as follows: low rates correspond to high thresholds, which means that the thresholded spikes are very sparse and there is thus less dependency among them. Therefore the difference between zero-th and first order entropy gets smaller. This means that at low rates it is harder to exploit the significance map correlation between neighboring samples, unless the *values* are correlated (but in the present example they are not: only their *positions* are).

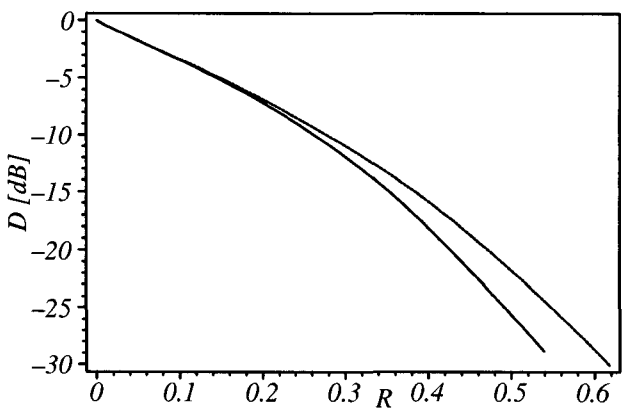

Figure 3.8: Upper bounds for memoryless BG spike with $p = 1/16$ (top curve) and first order Markov Gaussian spike with $\alpha = 1/30$ and $\beta = 1/2$, yielding same marginal probability $p = 1/16$ (bottom curve).

Discussion We proposed the spike process as a model for sparse transform coefficients. However, comparing with Figure 1.1 we see that its $D(R)$ behavior is very different from the one observed in actual image coders. The reason lies in the mixed nature of the model: at large rates, we have to spend the rate to code the discrete part[3], but in turn the distortion decay will be steeper, inversely proportional to the probability of nonzero samples. Conversely, by the tightness of the Shannon lower bound,

[3]If the pdf is continuous, for $t \to 0$ we need no side information rate, or $\lim_{t\to 0} h(\mu(t))$. Conversely, the spike case will top out at $h(p)$ (assuming $p < 1/2$).

a continuous random variable cannot have an asymptotic distortion decay other than the well known -6 dB per bit. We conclude that the spike process is a bad model for our intended application, or, to put it more positively, it is a solution in search of a problem.

3.4 Gaussian Mixture Model

As became clear in the above discussion, we should stick to continuous densities when modeling transform coefficients. One of the more common approaches to density estimation is based on Gaussian mixtures. In this section we will analyze a simple i.i.d. Gaussian mixture model, where a hidden binary memoryless source picks one of two zero mean memoryless Gaussian sources. This is a generalization of the spike model, where one source had zero variance. The model pdf is:

$$f(x) = pf(x|S=1) + (1-p)f(x|S=2) \tag{3.22}$$

where S is the hidden state selecting a source, and

$$f(x|S=i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-x^2/2\sigma_i^2}. \tag{3.23}$$

Such models have been used quite successfully in various applications, see [12] and references therein. To get realistic estimates for the parameters, we used a version of the EM algorithm [7] on the wavelet coefficients of the Lena image (transformed with SPIHT [42]). The scaling coefficients were discarded beforehand, since the model (3.22) does not fit them (due to nonzero mean); and therefore they are not accounted for in the rate distortion bound.

Figure 3.9 shows the magnitude bound and the empirical $D(R)$ computed with Blahut's algorithm [4]. Up to the knee, which is typical for image coding distortion rate curves, the distortion decays more rapidly. This means mainly that the sparse coefficients from the high variance source are retained by the thresholding operation. At higher rates, the coefficients from the low variance source also start being significant. *The empirical $D(R)$ curve crosses the bound for numerical precision reasons: the* pdf input to Blahut's algorithm should be more finely quantized around zero, where the low variance source is determinant. Another apparent fact is that at higher rates the bound becomes loose, because it converges to the trivial Gaussian upper bound $B(0,R)$. At high rates (very low threshold), $D(R)$ of a thresholded rv is close to $D(R)$ of the non-thresholded rv, which is far from the Gaussian upper bound for the mixture density in our example. A tighter upper bound, which holds also for densities other than Gaussian mixtures, will be presented in Section 3.5.

3.4.1 Lower Bound on $D(R)$ of Gaussian Mixtures

Since Gaussian mixtures are a popular tool to approximate unknown densities, it is useful to also have a lower bound on their rate distortion function. The Gaussian

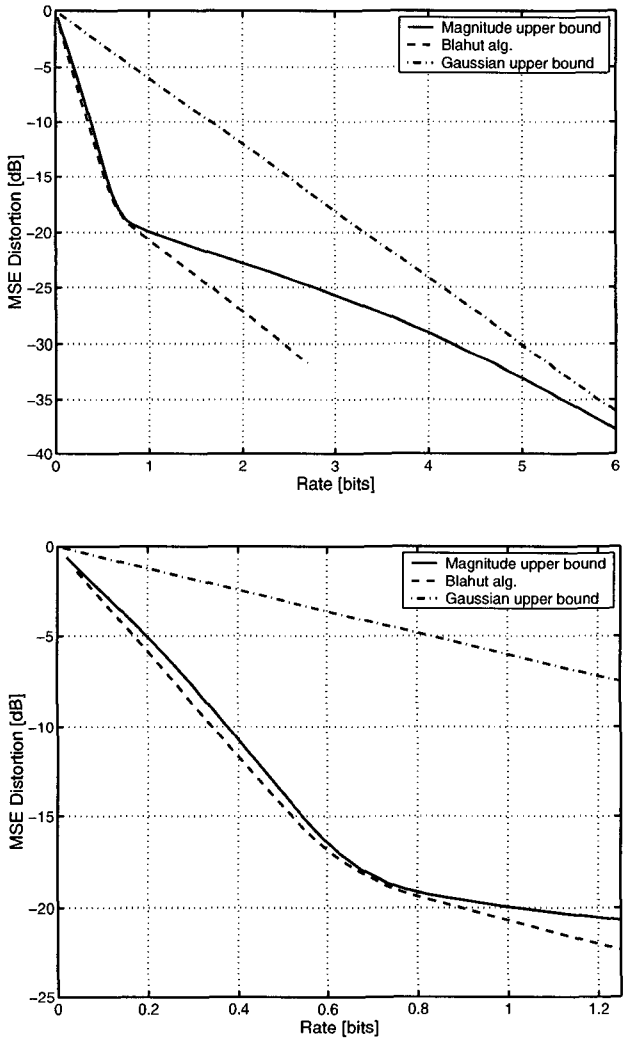

Figure 3.9: Gaussian mixture model for wavelet (detail) coefficients: upper bound on TOQ distortion rate function and empirical $D(R)$. The top curve is the trivial Gaussian upper bound $B(0, R)$. Model parameters, normalized to unit variance: $p = 0.9141$, $\sigma_1^2 = 0.01207$ and $\sigma_2^2 = 11.51$. At bottom, detail of low-rate region.

mixture source can be viewed as a finite alphabet discrete memoryless source S that switches between $|\mathcal{S}|$ Gaussian sources $\mathcal{N}(m_s, \sigma_s^2)$ with selection probabilities $w_s = Pr\{S = s\}$. A lower bound on $D(R)$ is found by assuming that an oracle provides the hidden state variable S to the source encoder. Since $S \to X \to \widehat{X}$ form a Markov chain, we have

$$I(X; \widehat{X}|S) \leq I(X; \widehat{X}). \tag{3.24}$$

We observe that $R_{lb}(D) := \min_{p(\hat{x}|x,s) \in Q_D} I(X; \widehat{X}|S)$ (with $Q_D = \{p(\hat{x}|x, s) : E(X - \widehat{X})^2 \leq D\}$) can be computed exactly by solving the following standard rate allocation problem (see Appendix A):

$$D_{lb}(R_{lb}) = \min_{\{R_s\}} \sum w_s \sigma_s^2 2^{-2R_s} \tag{3.25}$$

subject to

$$\sum w_s R_s = R_{lb} \text{ and } R_s \geq 0. \tag{3.26}$$

This yields the lower bound $D(R) \geq D_{lb}(R)$. We note that this bound can be seen as a special case of a conditional rate distortion function [22].

Figure 3.10 shows the lower bound, together with the upper bound to be presented in Section 3.5 and the (R, D) points achieved by a scalar bitplane quantizer (applied to $3 \cdot 10^5$ pseudo-random samples from mixture source; significance maps are entropy coded, sign and refinements bits left uncoded). At low rates, thresholding with simple scalar quantization performs very close to the R/D optimum.

3.4.2 Modeling Dependencies across Wavelet Scales

Practical wavelet coders try to exploit dependencies between coefficients within and across scales. One expects to find some correlation, because many popular filters are not perfectly orthogonal and also, by the partial non-randomness of images, they can never achieve total decorrelation. Data structures such as zero trees are used to track dependencies across scales, from coarse to fine. The in-scale dependencies are captured by a small causal context around the coefficient being processed. There are no good continuous models of this correlation (e.g. Gauss-Markov, AR), also because their existence would imply that we could have built a better decorrelating transform in the first place. Instead, practical schemes such as SPIHT [42] use the significance map to extract a context for the binary arithmetic encoder that encodes the significance of a coefficient. More elaborate schemes also consider the sign, but yield a relatively small gain, which can be improved if the signs are used in an estimate of coefficient magnitude [13]. This sums up to the general agreement that the most significant bit of the coefficients (represented in base 2) "carries" most of the correlation that can be exploited in practical compression schemes [50]. Over the last years, several authors

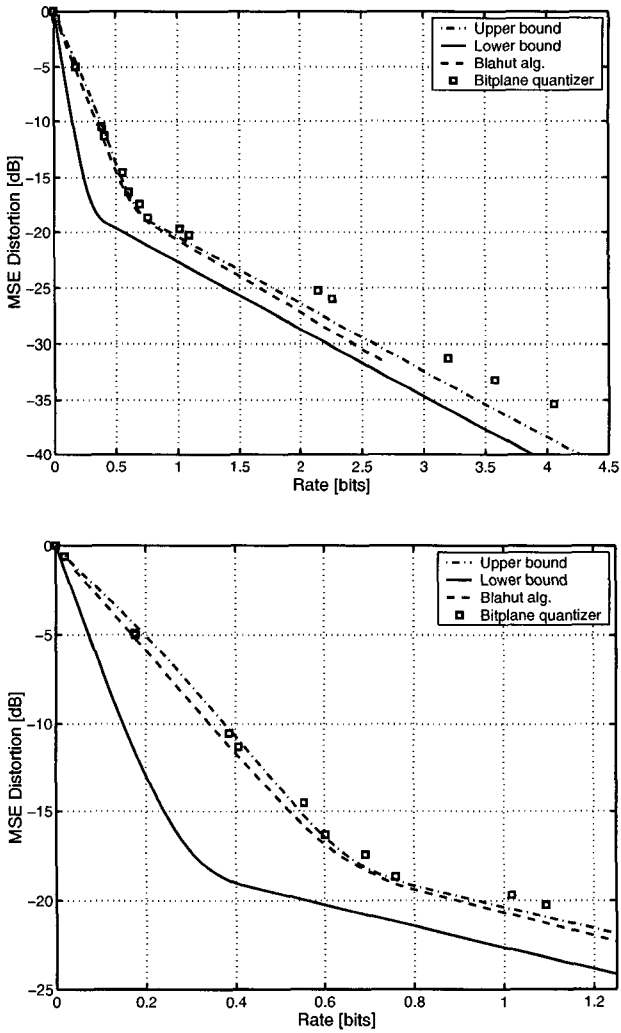

Figure 3.10: Gaussian mixture model for wavelet (detail) coefficients: Low- and high-rate upper and lower bounds on distortion rate. The middle curve is the empirical $D(R)$, the boxes denote (R, D) points achieved with a bitplane quantizer. At bottom, detail of low-rate region.

have also remarked that it is not necessary to use a tree data structure to catch dependencies across scales. It suffices to reorder coefficients so that dependent coefficients are close to each other in a linear arrangement. Alternatively, larger contexts containing all spatially neighboring coefficients are used in the arithmetic encoder, as is done for example in the JPEG 2000 standard. The main advantage compared to the tree approach is the absence of large global data structures, a disadvantage is the increased computational complexity (the arithmetic encoder has to encode lots of zeros that the tree would represent with one node) [34].

To summarize, we see that trying to find a good model of coefficient dependencies is a bit of a "catch-22". If the model is discrete (based on most significant bits), then the appropriate framework seems universal lossless coding, which doesn't deal with distortion. If the model is continuous we could hire an engineer to build a better decorrelating transform. Finally, we remark that the Markov spikes presented in Section 3.3.1 are certainly *not* a good model!

3.5 High-Rate Bound

The Gaussian mixture example made it obvious once more that the magnitude upper bound becomes loose at high rates, as it approaches the Gaussian upper bound. There is definitely an opportunity for a tighter high-rate bound. Besides that we need an exception to confirm the rule that this thesis focuses on low rates.

In Section 2.3 we bounded the rate distortion function of the significant samples above threshold with Eq. (2.15), since we were looking at thresholding oligoquantization. However, a better $D(R)$ bound may result if the samples below threshold are also considered for coding, i.e. upper bounded as Gaussians. The corresponding quantization method is a straightforward generalization of oligoquantization.

Definition 3.5 *Magnitude Classifying Quantization (MCQ) is a two step quantization procedure for i.i.d. sources. First, the samples are classified into significant and insignificant ones, according to whether their magnitude lies above or below a chosen threshold, respectively. The class of each sample is sent to the decoder as side information. Then each class is quantized separately using one of two (high-dimensional) quantizers, whose rates are chosen in order to minimize overall distortion.*

In the usual manner, we are going to state an upper bound on distortion rate which is also an upper bound on operational $\delta(R)$ of MCQ.

Theorem 3.3 (High-Rate Magnitude Bound) *Let the variances of the insignificant and the significant samples be*

$$\sigma_0^2(t) = \mathrm{E}[X^2 \,|\, |X| < t] = \tfrac{\sigma^2 - A(t)}{1-\mu(t)} \tag{3.27a}$$

and

$$\sigma_1^2(t) = \mathrm{E}[X^2 \,|\, |X| \geq t] = \tfrac{A(t)}{\mu(t)}, \tag{3.27b}$$

respectively. Then for all

$$R \geq R_{min}(t) = h(\mu(t)) + \tfrac{1}{2} \ln \tfrac{\sigma_1^2(t)}{\sigma_0^2(t)} \tag{3.28}$$

distortion rate of a memoryless source is upper bounded by

$$D(R) \leq B_{hr}(t, R) = c(t) e^{-2R}, \tag{3.29}$$

where

$$c(t) = \exp\left[3h(\mu(t)) + (1 - \mu(t)) \ln(\sigma^2 - A(t)) + \mu(t) \ln A(t)\right].$$

The best asymptotic upper bound for $R \to \infty$ is obtained by numerically searching the $t_0 \in [0, \infty)$ that minimizes $c(t)$. Since $\lim_{t \to +0} c(t) = \sigma^2$, the Gaussian upper bound is always a member of this family of bounds.

Proof: The thresholding process splits the input into two "sources" with variances given in (3.27a). A given sample belongs with probability $w_0 = 1 - \mu$ to the first source, and with probability $w_1 = \mu$ to the second (the dependence on t is dropped for convenience). We assign rate $R_{V,i}$ to source i. Then the average rate per sample is

$$R = h(\mu) + w_0 R_{V,0} + w_1 R_{V,1}$$

and the average distortion can be upper bounded with

$$D \leq w_0 \sigma_0^2 e^{-2R_{V,0}} + w_1 \sigma_1^2 e^{-2R_{V,1}}.$$

We see that optimizing the bound is a standard weighted rate allocation problem, except for the side information term $h(\mu)$ that gets added to the rate. The detailed solution can be found in Appendix A; it leads immediately to (3.29). Condition (3.28) guarantees that the rate allocation is proper, i.e. both $R_{V,i}$ are non-negative. □

Corollary 3.4 *The (low-rate) magnitude bound (2.18) and the high-rate magnitude bound (3.29) coincide in proper non-boundary local extrema of the high-rate bound, provided that $f(x) > 0$ over $supp(f)$. More formally,*

$$R^*(t) = R_{min}(t) \iff B(t, R^*(t)) = B_{hr}(t, R_{min}(t)) \iff c'(t) = 0. \tag{3.30}$$

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

$$\begin{aligned} c'(t) &= c(t) \left[\mu'(t) \ln \tfrac{\sigma_1^2(t)}{\sigma_0^2(t)} + A'(t) \left(\tfrac{1}{\sigma_1^2(t)} - \tfrac{1}{\sigma_0^2(t)}\right) + 2\mu'(t) h'(\mu(t))\right] \\ &= -2 f(t) c(t) \left[\ln \tfrac{\sigma_1^2(t)}{\sigma_0^2(t)} + t^2 \left(\tfrac{1}{\sigma_1^2(t)} - \tfrac{1}{\sigma_0^2(t)}\right) + 2h'(\mu(t))\right]. \end{aligned} \tag{3.31}$$

Since we assumed $f(t) > 0$ 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 let's look at the middle equation:

$$\begin{aligned} B(t, R^*(t)) &= A(t) \exp\left[2h'(\mu(t)) + \gamma(t) + W_{-1}(-\gamma(t)e^{-2h'(\mu(t))-\gamma(t)})\right] \\ &\quad + \sigma^2 - A(t) \\ &\stackrel{!}{=} B_{hr}(t, R_{min}(t)) = \sigma_0^2(t) = \tfrac{\sigma^2 - A(t)}{1-\mu(t)}, \end{aligned} \tag{3.32}$$

which after taking the logarithm is equivalent to

$$W_{-1}(-\gamma(t)e^{-2h'(\mu(t))-\gamma(t)}) = -2h'(\mu(t)) - \gamma(t) - \ln \tfrac{\sigma_1^2(t)}{\sigma_0^2(t)}. \tag{3.33}$$

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

$$\left(-2h'(\mu(t)) - \gamma(t) - \ln \tfrac{\sigma_1^2(t)}{\sigma_0^2(t)}\right) \tfrac{\sigma_1^2(t)}{\sigma_0^2(t)} e^{-2h'(\mu(t))-\gamma(t)} = -\gamma(t)e^{-2h'(\mu(t))-\gamma(t)}, \tag{3.34}$$

which is equivalent to

$$0 = \left[\ln \tfrac{\sigma_1^2(t)}{\sigma_0^2(t)} + 2h'(\mu(t)) + \gamma(t)(1 - \tfrac{\sigma_1^2(t)}{\sigma_0^2(t)})\right] e^{-2h'(\mu(t))-\gamma(t)}. \tag{3.35}$$

Observing that $\gamma(t) = \frac{t^2}{\sigma_1^2(t)}$ shows that the term in brackets is equal to the bracketed term (3.31), so that (3.35) implies $c'(t) = 0$. Finally, the leftmost expression in (3.30) is

$$\begin{aligned} 0 &= R^*(t) - R_{min}(t) \\ &= -\tfrac{1}{2}\mu(t)\left[2h'(\mu(t)) + \gamma(t) + W_{-1}(-\gamma(t)e^{-2h'(\mu(t))-\gamma(t)})\right] - \tfrac{1}{2}\mu(t) \ln \tfrac{\sigma_1^2(t)}{\sigma_0^2(t)}. \end{aligned} \tag{3.36}$$

Since t is not allowed to be on the right support boundary, we can divide (3.36) by $\mu(t) > 0$ and after rearranging terms we get exactly equation (3.33). Thus also the first condition is equivalent to $c'(t) = 0$. □

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 3.4 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 (2.18) is again tighter; it becomes looser up to $R_{min}(t_0)$, while from that rate on (3.29) will be the loosest bound. So far we have found no examples of densities that lead to multiple local extrema of $c(t)$.

Example 3.1 (Comparison with Gerrish-Schultheiss Upper Bound) Now that we have both a low-rate and a high-rate upper bound, a comparison with previously known bounds is in order. By its design, the high-rate bound will never be weaker than the Gaussian upper bound. Therefore we limit the comparison to the Gerrish-Schultheiss bound (equation 1.15) stated in the introduction [18]. As mentioned there, its main disadvantage is that in general it needs a double numeric integration — with the usual Gaussian exception. For the Gaussian mixtures analyzed in the preceding section, the marginal (codebook) density $q(y)$ (equation 1.16) will again be a Gaussian mixture. Hence only a single numeric integration is needed to compute the differential entropy $H(Y)$.

Figure 3.11 plots the upper bounds for the by now familiar Gaussian mixture example (modeling wavelet coefficients). The advantage of the magnitude classification bounds at low to medium rates is evident: they are tighter and they are simpler to compute than the Gerrish-Schultheiss bound. However, at higher rates the Gerrish-Schultheiss bound will necessarily win over the high-rate bound, since it converges to the Shannon lower bound (it is asymptotically tight).

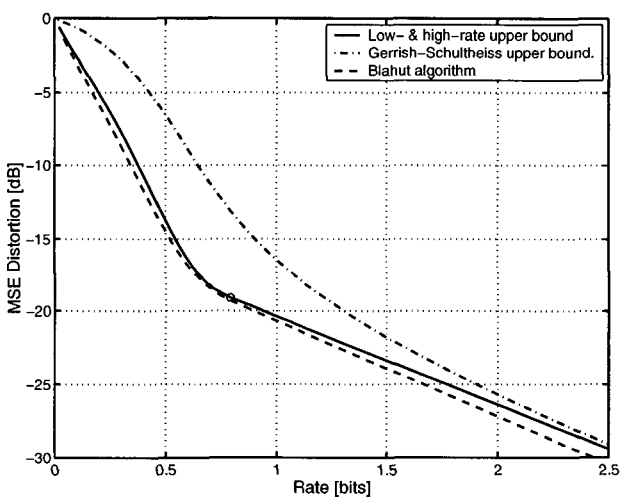

Figure 3.11: Gaussian mixture model for wavelet (detail) coefficients: Comparison with Gerrish-Schultheiss upper bound. The circle marks the switch-point from low-rate to high-rate magnitude bound.

3.5.1 An Upper Bound on Differential Entropy

We exploit the trivial fact that an upper bound to $D(R)$ is also an upper bound to the Shannon lower bound.

Corollary 3.5 *Let $\mu_0 = \mu(t_0)$ and $A_0 = A(t_0)$ be the quantities yielding the best asymptotic upper bound in Theorem 3.3. Define the probability mass functions*

$$\boldsymbol{\mu}_0 = [\mu_0, 1-\mu_0], \quad \boldsymbol{a}_0 = [\tfrac{A_0}{\sigma^2}, 1-\tfrac{A_0}{\sigma^2}].$$

If the underlying pdf $f(x)$ is absolutely continuous on $\mathbb{R}$, the differential entropy $H(X)$ can be upper bounded by

$$H(X) \leq \tfrac{1}{2}\ln(2\pi e\sigma^2) + h(\mu_0) - \tfrac{1}{2}D(\boldsymbol{\mu}_0\|\boldsymbol{a}_0). \tag{3.37}$$

Proof: By definition, $A(t)$ is non-negative, monotonically decreasing on $[0,\infty)$ with $A(0) = \sigma^2$. Therefore $\boldsymbol{a}_0$ is a proper pmf and thus the divergence in (3.37) is well defined. Since $f(x)$ is absolutely continuous, the Shannon lower bound is also well defined:

$$R(D) \geq R_{SLB}(D) = H(X) - \tfrac{1}{2}\ln(2\pi eD).$$

We may combine this with (3.29) expressed as a bound on $R(D)$:

$$R(D) \leq \tfrac{1}{2}\ln\tfrac{c(t_0)}{D} = \tfrac{3}{2}h(\mu_0) + \tfrac{1-\mu_0}{2}\ln(\sigma^2 - A_0) + \tfrac{\mu(t)}{2}\ln A_0 - \tfrac{1}{2}\ln D$$

to obtain

$$H(X) \leq \tfrac{1}{2}\ln(2\pi e\sigma^2) + \tfrac{3}{2}h(\mu_0) + \tfrac{1-\mu_0}{2}\ln(1-\tfrac{A_0}{\sigma^2}) + \tfrac{\mu_0}{2}\ln\tfrac{A_0}{\sigma^2}.$$

from which the corollary follows instantly. □

For $t_0 = 0$, that is $\mu_0 = 1$, the bound (3.37) reduces to the well known Gaussian upper bound on differential entropy. Of course we are more interested in random variables with very small (i.e. negative!) differential entropy, because these can be compressed more easily. In that case the divergence term must be very large, and the "side information" term $h(\mu_0)$ becomes negligible. Therefore low differential entropy is essentially due to a "variance distribution" $\boldsymbol{a}_0$ that is skewed compared to the probability distribution $\boldsymbol{\mu}_0$.

Example 3.2 Table 3.1 lists the upper bound and the entropy of a generalized Gaussian for several values of the shape parameter β. The bound is quite loose, compare also with Figure 2.6. For peaked Gaussian mixtures such as the one we studied in Section 3.4 the bound is much tighter.

Shape	Entropy	Upper bound
2 (Gaussian)	1.4189	1.4189
1 (Laplacian	1.3466	1.4064
1/2	0.9925	1.1624
1/8	-2.2430	-1.3100

Table 3.1: Differential entropy bounds for generalized Gaussian sources. (Unit variance, entropy in nats.)

3.5.2 Coding Gain Revisited

In the study of linear transform coding, the coding gain measures the compression gain of a transform coding system with quantizer bit allocation compared to a scalar quantizer system without transform. Here we show how the high-rate upper bound leads to an expression that is reminiscent of the coding gain of a two-dimensional transform coding system.

Let us quickly go through the derivation of the classical transform coding gain. Consider a real-valued, time-discrete, stationary and ergodic process $\{X_k\}$ with mean zero and variance σ^2. The samples are grouped into blocks $\boldsymbol{X} = [X_{i=1}^N]$ of length N and transformed with an orthonormal transform T:

$$\boldsymbol{Y} = T\boldsymbol{X}.$$

By Parseval's equality the quantization error in the signal domain will be equal to the error in the transform domain:

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

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 nonzero mean case. Let $\sigma_i^2 = \mathrm{E}\, Y_i^2$ be the variance of the i-th component, i.e. transform coefficient. Note that we actually mean a random variable when we talk about coefficients/components. If we use N scalar quantizers to quantize $\boldsymbol{Y}$, the optimal high-rate bit allocation is easily found using Lagrangian optimization (see Appendix A, with weights $w_i = 1/N$).[4] From Equation (A.8) we get an average distortion of the form $D = (\prod_{i=1}^N \sigma_i^2)^{1/N} e^{-2R}$. This can be compared with the distortion

[4]This uses the assumption that either the signal is a correlated Gaussian process (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 $D = ce^{-2R}$.

of a scalar quantizer applied to the X_i, which is $D = \sigma^2 e^{-2R} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\, X_i^2 e^{-2R}$. In fact, the *transform coding gain* can be defined as the ratio of the distortion of direct scalar quantization of the signal samples over scalar quantization of the transform coefficients (with bit allocation):

$$G_{TC} = \frac{\frac{1}{N}\sum_{i=1}^{N} \sigma_i^2}{\left(\prod_{i=1}^{N} \sigma_i^2\right)^{1/N}}. \tag{3.38}$$

In purely algebraic terms Equation (3.38) is the ratio of the arithmetic mean of the coefficient variances to the their geometric mean, which is often used as the "axiomatic" definition of coding gain. Our short derivation gives some additional insight on the implicit assumptions, namely high rate and (near-)Gaussianity.

Now it is obvious that we can define a measure of coding gain for magnitude classifying quantization by considering the ratio of the Gaussian upper bound to the high-rate upper bound (3.29).

Definition 3.6 *The coding gain for optimal*[5] *magnitude classifying quantization is*

$$G_{MCQ} = \frac{c(0)}{c(t_0)} = \frac{\sigma^2}{c(t_0)} = \frac{\mu(t_0)\,\sigma_1^2(t_0) + (1-\mu(t_0))\,\sigma_0^2(t_0)}{e^{2h(\mu(t_0))}\,\sigma_1^{2\mu(t_0)}(t_0)\,\sigma_0^{2(1-\mu(t_0))}(t_0)}, \tag{3.39}$$

where t_0 is the threshold yielding the tightest upper bound in Theorem 3.3.

Except for the additional side information term $e^{2h(\mu(t_0))}$, this definition corresponds to the classical coding gain (3.38) for two sources with weights $\mu(t_0)$ and $1-\mu(t_0)$. 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 also the MCQ coding gain will be large. At the same time the mixing approach obviously entails a loss in coding gain, that we study by means of an example.

Example 3.3 (Coding Gain Loss for Gaussian Mixtures) If the transform outputs zero mean Gaussian coefficients, where each has one of just two distinct variances, the resulting marginal density will be a two component Gaussian mixture like the one studied in Section 3.4. We get the largest classical coding gain if for every sample we know from which of the two sources it was emitted. That situation corresponds exactly to the "oracle" lower bound presented in Section 3.4.1, and the coding gain is simply the distance in dB to the Gaussian upper bound. The coding gain loss is the ratio of MCQ coding gain (3.39) to classical coding gain (3.38), or the distance in dB from the lower bound (3.25) to the high-rate upper bound (3.29):

$$\Delta_{CG} = \frac{e^{2h(\mu(t_0))}\,\sigma_1^{2\mu(t_0)}(t_0)\,\sigma_0^{2(1-\mu(t_0))}(t_0)}{\sigma_{m0}^{2(1-w_1)}\,\sigma_{m1}^{2w_1}}.$$

[5]Here *optimal* refers to the tightest upper bound of Theorem 3.3; directly optimizing a MCQ would yield tighter bounds, because significant and insignificant samples differ in $D(R)$ behavior.

Note that here $\sigma_0^2(t_0)$ denotes the variance of the sub-threshold samples, while σ_{m0}^2 is the first mixture variance. Figure 3.12 contains contour plots of the coding gain and the MCQ loss Δ_{CG} for different ratios $\theta^2 = \sigma_{m1}^2/\sigma_{m0}^2$ of the mixture variances and weights $w_1 = 1 - w_0$ ($\theta = 1$ is the Gaussian pdf). Large θ and small w_1 lead to peaked densities; for example, the wavelet coefficient mixture from Section 3.4 has $\theta \approx 30.9$ and $w_1 \approx 0.09$. 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 Figure 3.10.

The above definition of coding gain loss is based on the assumption that we are actually mixing two Gaussian sources with distinct variances (i.e. $\theta > 1$). What if we only have a single source with the same marginal mixture density? Then the lower bound is not achievable for $\theta > 1$ and thus a better definition of coding gain loss is the ratio of the high-rate upper bound to the Shannon lower bound:

$$\Delta_{CG(SLB)} = \frac{e^{2h(\mu(t_0))}\,\sigma_1^{2\mu(t_0)}(t_0)\,\sigma_0^{2(1-\mu(t_0))}(t_0)}{\exp[2H(X) - \ln(2\pi e)]}.$$

The differential entropy $H(X)$ has to be computed with numerical integration methods. Figure 3.13 plots the coding gain and the MCQ loss for this case. The MCQ loss is remarkably low over a wide range of parameter values. This shows that the magnitude classification quantization approach is very effective for such sources. Let us also remark that in this example the optimal MCQ threshold t_0 was always larger than the threshold for the maximum likelihood classification, $t_{ML} = \sqrt{\ln\theta^2/(1-\theta^{-2})}\sigma_{m0}$. This is quite natural, since the goal of the classification is a tight distortion bound, not the optimal distinction of the two component sources.

Gaussian mixture models have often been used in image compression, for example a classification approach has been proposed in [25]. The authors consider the joint numerical optimization of the classifier and (high-rate) uniform quantizers for each of the N classes. Their simulation results indicate that for typical image data $N = 2$ classes yield a substantial improvement over a single class. Adding more classes gives only minor gains over $N = 2$, which supports our observation that a two-component Gaussian mixture is a good basic model for wavelet coefficients.

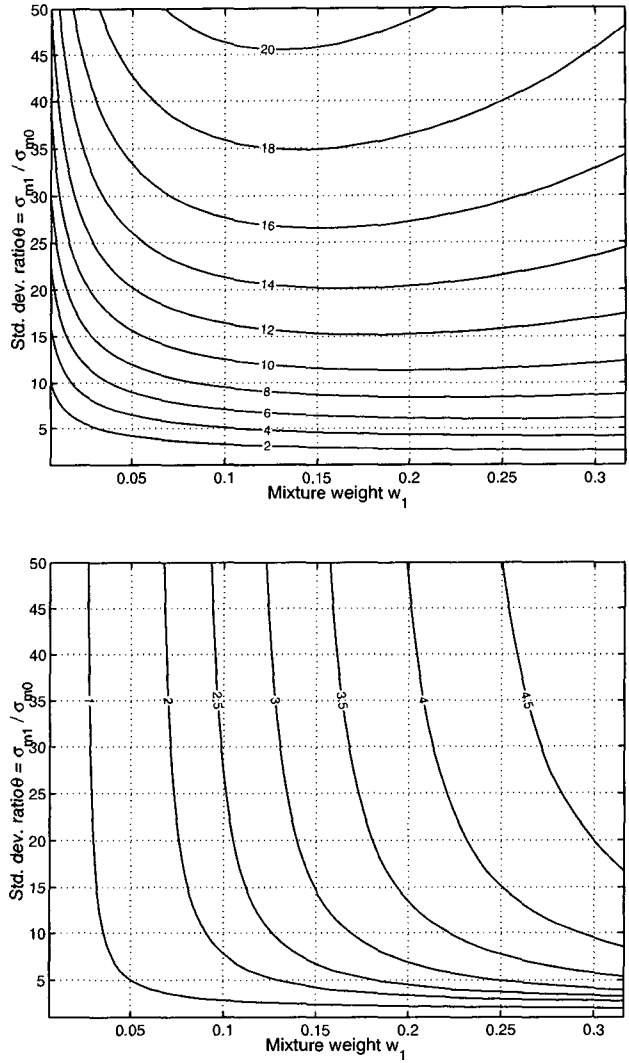

Figure 3.12: Magnitude classifying quantization (MCQ) of two component Gaussian mixtures (GM). *Top:* Coding gain G_{TC} for unmixed sources (equivalent to GM lower bound). *Bottom:* Coding gain loss for MCQ of the mixture.

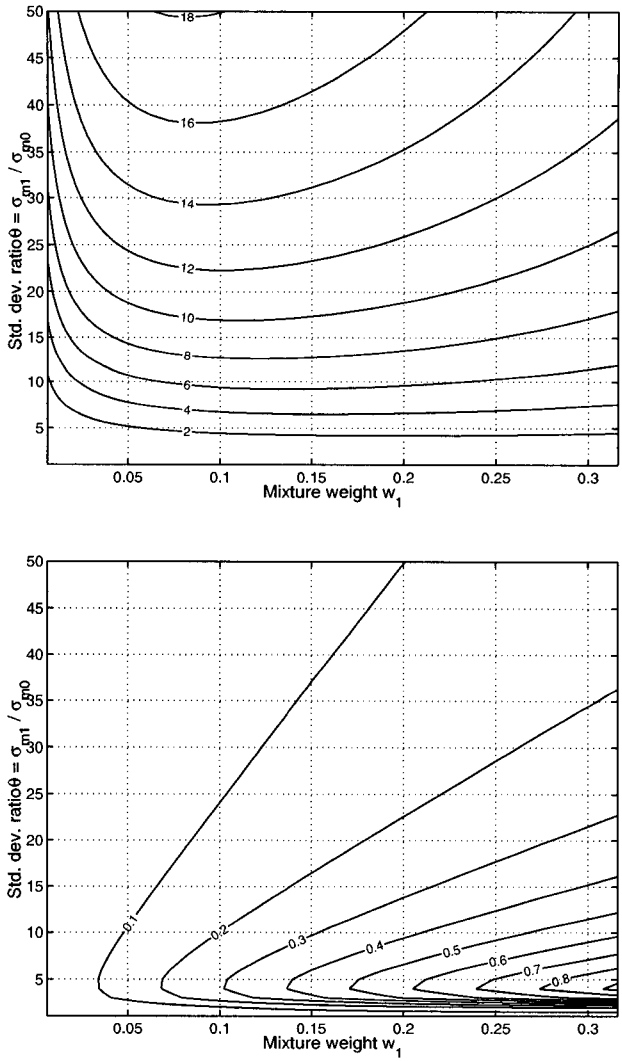

Figure 3.13: Magnitude classifying quantization (MCQ) of two component Gaussian mixtures (GM). *Top:* Coding gain G_{SLB} for mixture source (equivalent to Shannon lower bound). *Bottom:* Coding gain loss for MCQ of the mixture.

Chapter 4

Extensions

This chapter extends the results of the previous chapters in various directions (its self-explanatory working title was "odds and ends"). To get started, we outline a way to generalize the magnitude upper bounds to sources with arbitrary, asymmetric densities. The section thereafter deals with the key quantities $\mu(t)$ and $A(t)$, showing how they yield a succinct characterization of the rate distortion behavior of a random source. Also, we outline how the bounds can be estimated directly from a set of samples, without having to estimate the underlying probability density. The core of the chapter addresses the extension of oligoquantization to higher dimensions. Two examples will demonstrate that some improvements compared to scalar thresholding are possible. On the other hand we will show that asymptotically for large dimensions, oligoquantization amounts to simple multiplexing between no quantizer (rate zero) and whatever quantizer is used for the significant samples. This means that oligoquantization should only be considered at low rates *and* low dimensions.

4.1 Sources with Arbitrary Densities

In Chapter 2 we made the assumption that the random variable X has a symmetric probability density function, thus simplifying many calculations. This section will give a short outline of which results still hold when this condition is relaxed.

For an arbitrary pdf $f(x)$, we define the moment functions to be

$$\mu(t) = \int_{-\infty}^{-t} f(x)\,\mathrm{d}x + \int_{t}^{\infty} f(x)\,\mathrm{d}x, \tag{4.1a}$$

$$A(t) = \int_{-\infty}^{-t} x^2 f(x)\,\mathrm{d}x + \int_{t}^{\infty} x^2 f(x)\,\mathrm{d}x. \tag{4.1b}$$

All our upper bounds are based on the Gaussian upper bound, $D(R) \leq \sigma^2 e^{-2R}$. Therefore if the variance of the (in)significant samples is overestimated, the bounds

will still hold. Luckily this is precisely the case, since

$$\mathrm{Var}(X^2 \,|\, |X| < t) \leq \mathrm{E}[X^2 \,|\, |X| < t] = \frac{\sigma^2 - A(t)}{1 - \mu(t)}, \tag{4.2a}$$

$$\mathrm{Var}(X^2 \,|\, |X| \geq t) \leq \mathrm{E}[X^2 \,|\, |X| \geq t] = \frac{A(t)}{\mu(t)}. \tag{4.2b}$$

The other critical points in the proofs are the derivatives of the moment functions, their ratio should be $A'(t)/\mu'(t) = t^2$ (see the Proof of Theorem 2.3, at end). From (4.1) we see that the derivatives are

$$\mu'(t) = -f(-t) - f(t), \tag{4.3a}$$

$$A'(t) = [-f(-t) - f(t)]t^2, \tag{4.3b}$$

hence their ratio has indeed the desired value. If the pdf has finite support, additionally we have to make sure that at least one of $f(-t)$ and $f(t)$ is positive.

From Equations (4.2) and (4.3) we can conclude that the low-rate magnitude bound (Theorem 2.3) and its high-rate cousin (Theorem 3.3) hold for arbitrary source densities if the definitions (4.1) are used. The latter implies that also the entropy bound (Corollary 3.5) is valid. Notice that the bounds can by optimized by shifting the pdf: intuitively the mode of the density should be close to the origin, such that the distortion floor $\mathrm{E}[X^2 \,|\, |X| < t]$ is small (see also Section 4.3.1).

The case of the magnitude+sign bound (Theorem 2.4) is much more delicate. Basically we split the source into three sources (classes) with variances $\sigma_{-1}^2, \sigma_0^2, \sigma_{+1}^2$ and weights (probabilities) $\mu_{-1}, \mu_0, \mu_{+1}$. The high-rate bound can be computed in straightforward fashion, since it is simply a rate allocation problem for three sources. Conversely, things get complicated as soon as one or more of the three partial rates $R_{V,i}$ is zero. Of course we could force this back into the symmetrical pdf case by: i) Using always 1 bit to encode the sign, even if $\mu_{-1} \neq \mu_{+1}$, ii) Setting $R_{V,-1} = R_{V,-1}$ and iii) Upper bounding σ_{-1}^2 and σ_{+1}^2 by their maximum. However it is likely that this triple weakening will not yield any interesting bounds, therefore we do not explore it any further.

4.2 One graph says it all

Information theorists have a weakness for single-letter descriptions, which describe the essential characteristics of a problem with a single quantity. Unfortunately concerning rate distortion behavior there is not much that can be described by a single quantity: the variance of a random variable yields the Gaussian upper bound, and together with its entropy one can compute the Shannon lower bound. However, we need more information to get a more detailed picture, in particular at low rates. The bounds presented in the previous Chapters are quite demanding in that respect, since they require the knowledge of the (symmetric) pdf to compute the zero-th and second moments of the thresholded random variable. However in practical situations, where

one is given a sequence of samples from an unknown source, it can be quite tricky to obtain a reasonable estimate of the underlying density. On the other hand, sample moments are easy to compute and are statistically robust if the i.i.d. assumption holds.

In this section we will argue that the "moment profile" $A(\mu)$ is all we need to compute our bounds. Of course this is far from a single-letter characterization, but at least all relevant quantities can be read off directly from the graph of $A(\mu)$, whereas starting from the pdf $f(x)$ involves more computations. And the moment profile can be directly computed by ordering the (squared) sequence of samples.

Looking at the different magnitude bounds, we see that all quantities that are ever needed are t^2, $\mu(t)$ and $A(t)$. The latter two are obviously given by $A(\mu)$, while for the former we observe that

$$A'(\mu(t)) = \frac{\dot{A}(t)}{\dot{\mu}(t)} = \frac{-2f(t)t^2}{-2f(t)} = t^2, \tag{4.4}$$

where prime denotes a derivative with respect to μ, and dot w.r.t. t. So the slope of the moment profile $A(\mu)$ is exactly the squared threshold t^2 we were looking for. If furthermore we assume that $f(x) > 0$ for all $t < t_{max}$ (the support of $f(x)$ has no holes, i.e. X takes on values anywhere in its range $(-t_{max}, t_{max})$), then we have

$$\begin{aligned} A''(\mu(t)) &= \frac{\dot{\mu}(t)\ddot{A}(t) - \dot{A}(t)\ddot{\mu}(t)}{\dot{\mu}^3(t)} = \frac{-2f(t)[-4f(t)t - 2f'(t)t^2] - 4f(t)t^2 f'(t)}{-8f^3(t)} \\ &= -\frac{t}{f(t)} < 0 \quad \forall t : 0 < t < t_{max}. \end{aligned} \tag{4.5}$$

We summarize the properties of $A(\mu)$:

- $A(0) = 0$ and $A'(0) = t_{max}^2$,
- $A(1) = \sigma^2$ and $A'(1) = 0$,
- $A(\mu)$ is strictly monotonically increasing on $(0, 1)$ by (4.4) and
- $A(\mu)$ is strictly concave-$\cap$ on $(0, 1)$ by (4.5).

An immediate consequence of the above is that the moment profile of a random variable X with finite range $(-t_{max}, t_{max})$ cannot reach points (μ, A) that lie above the tangent through $(0, 0)$ with slope t_{max}^2. However, by the high-rate bound (Theorem 3.3), or equivalently the entropy bound (Corollary 3.5), to get high compression ratios we would like a moment profile $A(\mu)$ that is as close as possible to the upper left hand corner. See also Figure 4.1, which is a contour plot of the differential entropy upper bound. Does this mean that a bounded random variable cannot be compressed as well as an unbounded one? We conjecture this is not the case, but rather it means that the underlying Gaussian upper bound does a bad job on bounded random variables.

Example 4.1 We take again the wavelet transform coefficients from the Lena image that we used in Section 3.4 to compute a Gaussian mixture model. However this time we take all of them, after subtracting the mean from the lowpass (scaling) coefficients (this is also done in SPIHT and costs negligible extra rate). The estimated moment

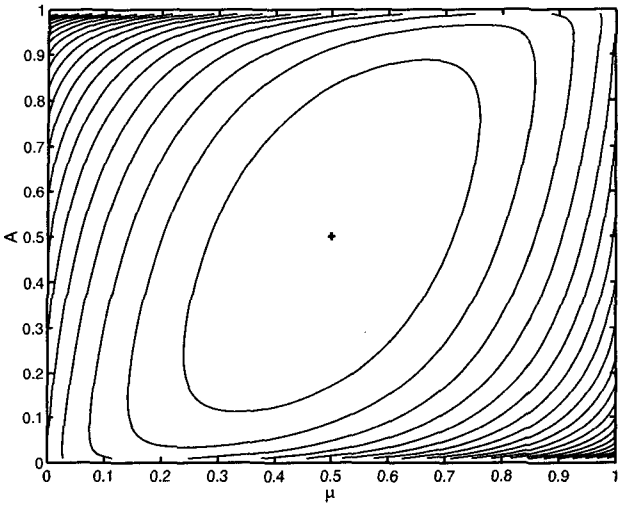

Figure 4.1: Contour plot of the differential entropy upper bound Eq. 3.37 in the (A, μ) plane (normalized to $A(1) = \sigma^2 = 1$, maximum at $(\frac{1}{2}, \frac{1}{2})$).

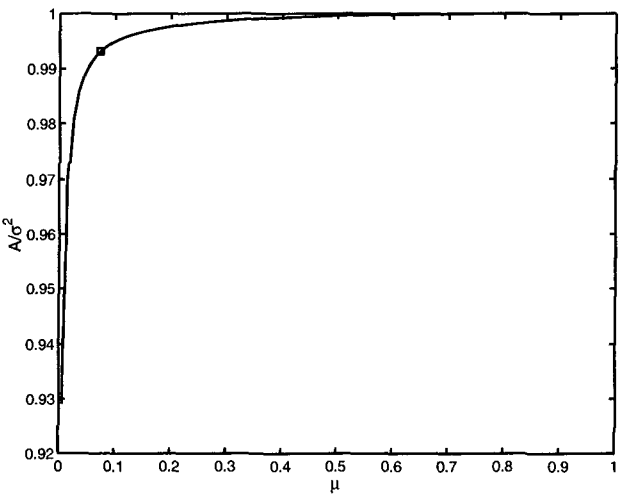

Figure 4.2: Empirical moment profile $A(\mu)$ for the Lena wavelet coefficients. The square marks the point (μ_0, A_0), where the high-rate bound is tightest.

profile $\hat{A}(\mu)$ is shown in Figure 4.2. We could fit a curve (e.g. a spline) to estimate the slope t^2, but actually there is no need to do so. Since we obtain $\hat{A}(\mu)$ from the ordered sample magnitudes $x_1 \geq x_2 \geq \ldots \geq x_k \geq \ldots \geq x_n$, for $\mu = k/n$ a reasonable estimate for t is $\hat{t} = x_k$, or maybe $\hat{t} = (x_k + x_{k+1})/2$. The reader might frown at this ad hoc approach, but since t is only needed to compute the "optimal" rate R^*, we always get a valid upper bound as long as $\hat{A}$ is a good estimate. This follows from the same reasoning which allowed us to loosen the bound in Section 2.5, and from the fact that μ is not estimated.

Figure 4.3 shows the resulting empirical magnitude upper bound and its high-rate continuation in comparison with a simple bit plane encoder (the same as in Section 3.4) and the SPIHT algorithm. Compared to the Gaussian mixture approach this empirical bound is certainly a better match to actual compression behavior. The gap between bound and SPIHT is mainly due to residual correlation among the coefficients, which SPIHT exploits via zero tree and conditional arithmetic coding. That also the scalar bit plane method beats the bound is mainly due to the Gaussian upper bound's looseness for high thresholds. Finally, at higher rates (above 2.5 bits) things look anomalous because the underlying image source is actually discrete. In fact both SPIHT and the bit plane method reach zero distortion at a rate of around 5.6 bits.

Comparing this example with the Gaussian mixture in Section 3.4, we suspect that the bound works best for samples from a real valued source with infinite support. These are of course exactly the assumptions implicitly made by using the Gaussian upper bound. In a sense the magnitude bound suffers from the inverse problem of Blahut's algorithm [4], which has to map infinite densities to a (implicitly bounded) finite set of source letters on the real line.

To assess the reliability of the estimated upper bound we observe that the magnitude bound has the form

$$A(\mu)[e^{\phi(\mu,t)} - 1] + \sigma^2.$$

Therefore the total estimation error depends only on the errors from estimating $A(\mu)$ and $\sigma^2 = A(\mu = 1)$. Since these are simple variance estimates, they can be easily analyzed with standard techniques.

Getting back to the "single letter quantity" question that prompted this section, we can argue that the best pick would be the *two* quantities μ_0 and A_0. These completely determine the high-rate upper bound from Theorem 3.3 and also its Corollary, the entropy upper bound. From the latter we can deduce that $A(\mu)$ has to be tangent to a contour in Figure 4.1, which allows us to gain a very rough picture of $A(\mu)$. This improves a bit if additionally we know whether the source is bounded or not (slope at $\mu = 0$). Of course from a guesstimate of $A(\mu)$ in the range $\mu = 0 \ldots \mu_0$ we could actually plot the low-rate bound.

Summing up, the magnitude bound provides a complement and/or an alternative to the Blahut-Arimoto algorithm (BA) for estimating rate distortion functions from samples of a memoryless source. Oftentimes, the BA does not provide reliable estimates (at very low and very high rates), because it is very hard to find the "right" quantization of the input density. To illustrate this point, we mention that we were

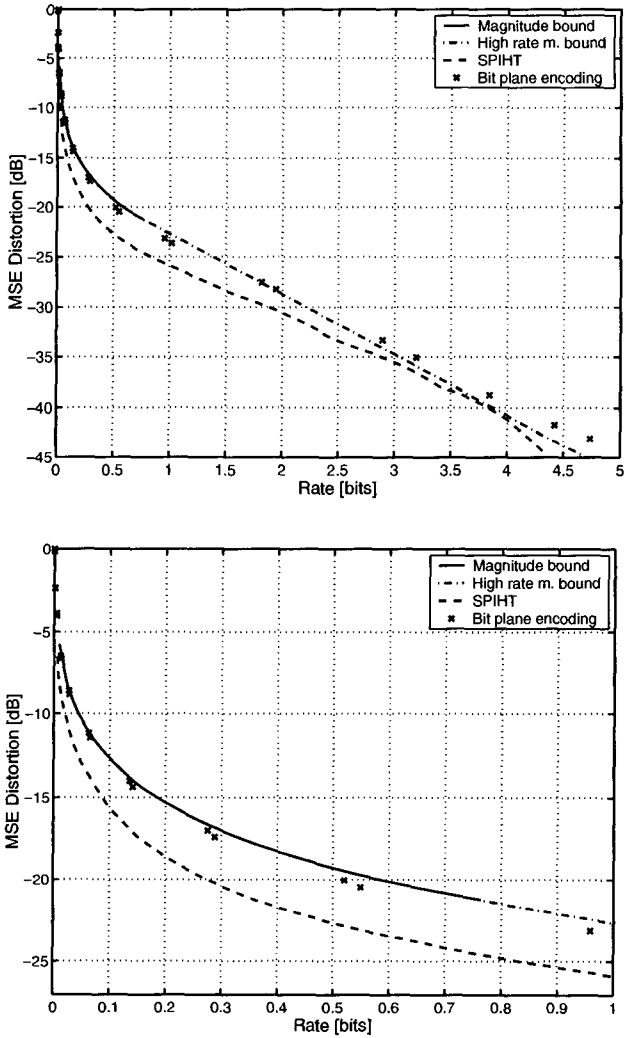

Figure 4.3: Empirical $D(R)$ bound computed from wavelet coefficients: Comparison with simple bit plane encoder and SPIHT image compression (bottom: low rate detail).

unable to get a decent $D(R)$ estimate for the above example.[1] Finally, we remark that it is also possible to estimate the M&S bound from the sample moment profiles $\hat{A}(\hat{\mu})$ and $\hat{\xi}(\hat{\mu})$, though often it doesn't improve on the magnitude bound (in the example it was looser for rates above 0.01 bits).

4.3 Going Multidimensional

In most coding problems considered in information theory, increasing the block size (the dimension) improves the code performance. Often this is termed the dimensional coding gain. Unfortunately, oligoquantization is an exception to this rule, it actually is most efficient in low dimensions. In this section we explore why this is so. Our approach will be based on multidimensional quantization, but the conclusions carry over to rate distortion bounds in the same spirit.

In the derivations of our bounds we assumed that the significant samples are encoded with an R/D optimal codebook, corresponding to an infinite-dimensional vector quantizer (even though in general scalar quantization suffices to beat the Gaussian upper bound). So the multidimensional generalization of thresholding oligoquantization does not concern the quantization step, but only the thresholding step. Before we extend scalar thresholding oligoquantization to vector thresholding, we want to restate once more what precisely makes it different from other quantization methods.

4.3.1 The Spirit of Oligoquantization

Let us recall the magnitude bound:

$$\begin{aligned} D(R) \le \Pr\{|X| \ge T\} \operatorname{Var}(X||X| \ge T) e^{-2R_V(T,R)} \\ + \Pr\{|X| < T\} \operatorname{Var}(X||X| < T). \end{aligned}$$

Clearly, the second term is a distortion floor which is independent of the rate per significant sample R_V. Now think of a M level scalar thresholding oligoquantizer and let $p_0 = \Pr\{X \in S_0\}$ be the probability of the zero bin (here $S_0 = \{x : |x| < T\}$). The entropy of the quantizer index $Y = Q(X)$ can be upper bounded as follows:

$$\begin{aligned} H(Y) = -\sum_{i=0}^{M-1} p_i \log p_i &\le -p_0 \log p_0 - (1-p_0) \log \frac{1-p_0}{M-1} \\ &= h(p_0) + (1-p_0)\log(M-1), \end{aligned} \tag{4.6}$$

since the conditional entropy $H(Y|Y \neq 0)$ is maximized by having uniform probability on the remaining $M-1$ indices. As a first attempt in building a *low rate* quantizer

[1] As already noted, the main problems of Blahut's algorithm at low rates stem from the discrete nature of the optimal reproduction alphabet. Both Fix [17] and Rose [38] proposed algorithms that alternatively optimize the reproduction alphabet and the conditional pmf's minimizing mutual information. Convergence properties for these schemes are harder to establish.

we would therefore try to make p_0 as large as possible, because this makes the bound (4.6) small.[2] At the same time we should keep the distortion incurred in the zero bin S_0 as small as possible, since it determines the distortion floor. Of course this is exactly what we are doing when we are thresholding a unimodal density which has its mode at the origin.

These principles carry over to vector quantizers with M codewords, where again we will try to identify a high probability/low distortion cell that keeps the rate low, without unduly increasing the distortion floor. Note that the "standard" way to build low-rate vector quantizers is to increase the dimension n until it is possible to place $M = e^{nR}$ reconstruction points that each have approximately the same probability and distortion contribution. The rate-distortion theoretic view of this is that the points should be uniformly spread over the typical set. Informally, classical VQ tends to symmetry, whereas the principle of oligoquantization is asymmetric.

The gist of oligoquantization is: given σ_0^2, find the set S_0 with the largest possible probability $p_0 = \Pr\{X \in S_0\}$ subject to the constraint $\mathrm{Var}(X|X \in S_0) \leq \sigma_0^2$. Or, for given p, find the set with probability at least p that has the smallest variance $\mathrm{Var}(X|X \in S_0)$. We will call this alternatively the high probability or the low variance set.

4.3.2 The Shape of the Low Variance Set

Finding the n-dimensional high probability/low variance set S_0 for an arbitrary pdf $f(x)$ is a challenging problem. It resembles a classical isoperimetric problem that can be attacked with variational methods, but it is more difficult since also the centroid is "variable".

The problem is simplified a bit if the density $f(x)$ is symmetric and strictly unimodal, which implies that $f(x)$ is maximal at $x = 0$ and that it is a decreasing function of $|x|$. Then also the k-dimensional product density $f(\boldsymbol{x}) = \prod_{i=1}^{k} f(x_i)$ will be decreasing along any ray starting at the origin. These properties enable us to give an *informal* recipe for building the high probability/low variance set:

- For vanishingly small volume $Vol(S_0)$, the pdf $f(\boldsymbol{x})$ can be assumed constant over S_0. Then the variance is only a function of the volume. We conclude that in this case, the set S_0 contains the origin (the mode), since by that it will have the largest possible probability for the given volume.

- Thanks to the point symmetry $f(-\boldsymbol{x}) = f(\boldsymbol{x})$, larger volume sets can be "grown" by adding pairs of points which are symmetric to the origin. Thus S_0 is point symmetric around the origin, which will also be the centroid, and hence the variance is minimal.

- The set S_0 contains no "holes", all points on a segment from the origin to a point

[2] And it minimizes the proportion $1 - p_0$ of samples that have to be quantized to one of the other $M - 1$ levels. Hence the name *oligo*quantization.

on the boundary ∂S_0 belong to S_0. Otherwise we could reduce the variance by "filling the holes", since $f(\boldsymbol{x})$ is decreasing along any ray starting at the origin.

The above arguments are not sufficient to determine the shape of the set S_0. For that task, we take a very practical approach and consider a block of n samples, where each sample is a k-dimensional vector. In the limit of large block lengths, S_0 will contain about $n_0 \approx p_0 n$ samples, and its sample variance will be close to σ_0^2. It is evident that for given n_0 we get the lowest variance set by taking the n_0 samples with smallest norm $\|\boldsymbol{x}_i\|^2$ (their mean will be $\approx \mathbf{0}$, thus $\mathrm{Var}(S_0) \approx \mathrm{E}_{S_0} \|\boldsymbol{X}\|^2$). From this we see that S_0 is a k-dimensional hyperball.

4.4 Circular and Spherical Thresholding

We present practical examples of two- and three-dimensional oligoquantizers for spherically symmetric densities such as the Gaussian. The thresholding and the quantizer will have the same dimension, as was the case for scalar Gaussian quantization in example 2.1. Therefore the two step oligoquantizer can again be seen as an ordinary k-dimensional quantizer in disguise.

By the discussion in the previous Section it is clear that we will classify as insignificant all samples within a circle of radius t centered at the origin (a sphere in 3D). At the decoder they will be mapped to the all zero vector. In two dimensions, the significant samples are quantized using a polar quantizer, whose $M-1$ reconstruction points will be regularly spaced on a circle of radius $t_1 > t$ that can be determined by the centroid rule. In three dimensions, the simplest approach is to use the vertices of a regular polytope [11] as reconstruction points. The general idea is to threshold with a k-dimensional sphere and then use a k-dim. *geometrically uniform* quantizer whose reconstruction (code) points lie on a larger sphere. We get the code points of such a geometrically uniform quantizer by letting a finite subgroup of $SL_k(\mathbb{R})$ act on a start vector.[3] The term "geometrically uniform" stems from the fact that all Voronoi cells will have the same shape. This approach with two concentric spheres (one is the threshold, the other contains the quantizer codebook) is clearly modeled after the one dimensional example, with one center bin and two outside bins. The main difference from other low-rate vector quantizers is that we impose the spherical center cell $\boldsymbol{x} : \|\boldsymbol{x}\|^2 < t^2$ (the loss due to spherical thresholding instead of using the Voronoi cells is negligible). Compare the scalar case were we prescribed the zero bin $[-t, t]$.

For a number of such quantizers we ran simulations with a Gaussian source. By varying the radius t different rates are achieved. The results are plotted in Figure 4.4 (2D) and Figure 4.5 (3D). The peculiar look of the plots comes from plotting the *differences* to the Gaussian distortion rate function, $D(R) = \sigma^2 2^{-2R}$. This makes their visual comparison easier. It comes as no big surprise that in two dimensions the best quantizer is hexagonal (plus the origin). Its gain compared to a three level scalar

[3] Think of a finite group built from reflections and rotations.

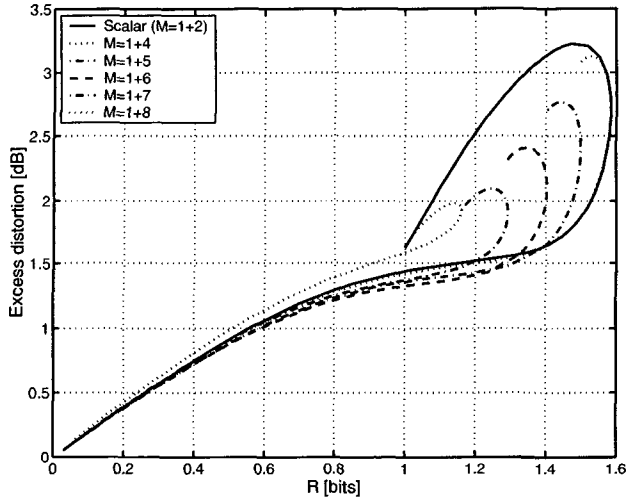

Figure 4.4: Circular thresholding plus polar quantization of Gaussian.

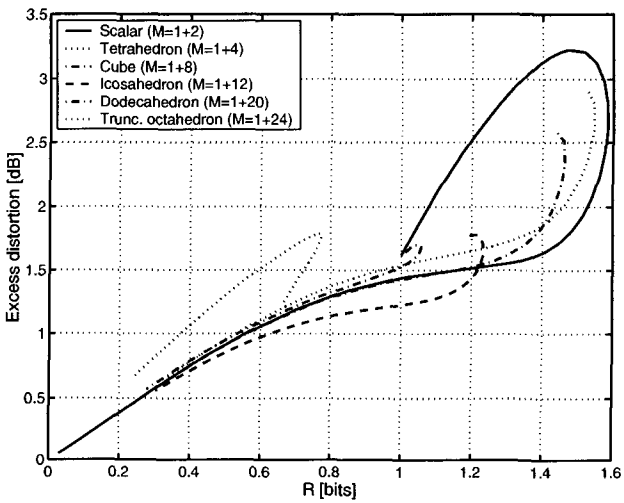

Figure 4.5: Spherical thresholding plus spherical quantization of Gaussian.

quantizer is about 0.1 dB at a rate of 1 bit. In three dimensions, the winner so far[4] is the icosahedron, which gains 0.2 dB at 1 bit.

Design and analysis of spherically symmetric quantization constellations, like those used for the significant samples, have been studied by several researchers. Polar quantization of Gaussian random variables has been studied by Bucklew and Gallagher in [5]; they also treated the non-Gaussian case in [6]. Spherical quantization in higher dimensions has been analyzed by Swaszek and Thomas [46]. There is also a close resemblance between multidimensional oligoquantization and root lattice vector quantizers [40]. The main difference is that oligoquantization starts by specifying an n-dimensional region (e.g a ball) containing the insignificant samples, after which the quantizer for the significant samples (outside this region) is designed. In general this region, which we also dubbed the low variance set, will not be a Voronoi cell of the quantizer constellation. Such a quantizer is thus inherently suboptimal, except in a few special cases. On the other hand, the rate of root lattice VQ can be varied by scaling the lattice and thus the corresponding Voronoi regions, i.e. the quantizer cells. However, the constellation points will not coincide with the centroids of the corresponding cells (again with rare exceptions). Hence also that scheme is suboptimal; whether oligoquantization or root lattice VQ is better has to be analyzed on a case-by-case basis.

4.4.1 High-Dimensional Dead End

The last theorem in this thesis states that it is not worth the trouble to extend oligoquantization to higher dimensions.

Theorem 4.1 *For large enough k, the distortion rate curve of k-dimensional oligoquantization is arbitrarily close to a line which corresponds to multiplexing (time sharing) between the trivial quantizer $Q(\boldsymbol{x}) = 0$ and some k-dimensional vector quantizer Q'.*

Proof: The only fact about typical sets that we need is that $\Pr\{\boldsymbol{x} \in \mathcal{T}_\epsilon^{(k)}\} > 1 - \epsilon$ for k sufficiently large. Now suppose we are growing our high probability/low variance set of probability mass $1 - p$. No matter what its shape is (!), at most a subset of arbitrarily small[5] probability mass ϵ lies *outside* the typical set. Therefore we are actually quantizing a proportion p of the typical set. This can be done for example using a vector quantizer for the original source with $M = 2^{kR_0}$ codewords (vectors) that achieves a distortion δ_0. Assuming that the codewords are evenly spread over the typical set, the high probability set will contain about pM of these codewords, which will be used each with equal probability $\frac{1}{pM}$. The rate is

$$R(p) \approx \tfrac{1}{k}[h(p) + p\log(pM)] \tag{4.7}$$

$$= -\tfrac{1}{k}(1-p)\log(1-p) + pR_0 \tag{4.8}$$

[4] We only tested the platonic solids and some other geometrically uniform constellations.

[5] Actually, choosing ϵ determines the minimal k for which the theorem holds.

and the distortion

$$\delta(p) = (1-p)\sigma^2 + p\delta_0. \tag{4.9}$$

Since $(1-p)\log(1-p)$ is bounded, the corresponding term in (4.7) vanishes for large k. Therefore high-dimensional thresholding followed by quantization is equivalent to multiplexing between $(R,\delta) = (0,\sigma^2)$ and $(R,\delta) = (R_0,\delta_0)$. □

Informally, we could characterize the first step of oligoquantization as asking the question "should we quantize this block of k samples?" There are two costs involved: the side information rate for giving the answer, and the distortion incurred if the answer is "no". The procedure pays off if the answer is most often "no" and the associated distortion is small. However, when k is large the theory of typical sequences tells us that the correct answer is almost always "yes". So there is no point in enforcing a "no", since then we will almost surely pay the maximal distortion cost σ^2.

The positive consequence of this theorem is that we can be content with the performance that low-dimensional oligoquantization delivers. If that's not enough, we have to shop for a true vector quantizer.

Chapter 5

Conclusion

The main contributions of this thesis can be summarized as follows:

- The distortion rate bounds presented in Chapter 2 are a useful tool to study the low-rate behavior of the peaked unimodal densities, like those that appear in modern transform coding systems. The given examples demonstrate that practical scalar oligoquantizers are often sufficient to even beat these bounds.
- Two-component Gaussian mixtures are shown to be an adequate model to catch the basic rate distortion behavior of wavelet transform coders.
- The high-rate bound of Section 3.5 allows us to introduce the notion of coding gain also for transform codes that are not based on the linear (KLT) paradigm.
- The sample-based bound estimates of Section 4.2 are a welcome complement to the Blahut algorithm, since the latter is only usable at medium rates.
- Finally, Chapter 4 explains why oligoquantization is most effective in low dimensions. This is good news to the users of scalar oligoquantization, but bad news to those who can afford the complexity of higher-dimensional systems.

The main weakness of our bounds is that we have no analytic tools to assess how tight they are compared to the actual rate distortion function.[1] On the other hand they can still serve as a benchmark for lossy source coders, in the sense that any decent coder should beat them!

Perhaps a deeper connection with rate distortion theory can be found by linking quantization theory with the results of Fix and Rose, which were mentioned in the introduction [17, 38]. Namely the fact that a low-rate optimal codebook will only need a few discrete reconstruction points reminds one of a scalar quantizer. However

[1] For an individual source we could of course compare with a numerical evaluation of the Shannon lower bound and/or determine $D(R)$ with Blahut's algorithm. However, we are more interested in the tightness for a *class* of sources.

we can't just compare with a scalar quantizer that has the same reconstruction points, because the rate would be much larger (even though the distortion would be slightly reduced thanks to the nearest neighbor rule).

Another point that merits further investigation is the tightness of the Gaussian upper bound for thresholded random variables. There are two possible goals: one is to tighten that bound and therefore also the upper bound on operational $\delta(R)$ of thresholding oligoquantization. That approach would be in the spirit of the magnitude plus sign bound of Section 2.4, but without using the first moment $\xi(t)$. The other possible goal is to upper bound the distortion of entropy coded scalar quantization of thresholded random variables, that is to generalize the Gish-Pierce asymptote. This would give us an asymptotically achievable upper bound on the operational rdf of scalar oligoquantizers.

To end this thesis on a humorous note, we could say that "an extensive list of related open problems can be requested from the author". This is one reason that makes research enjoyable, but also frustrating at times: it raises more questions than it ever answers.

Appendix A

Weighted Rate Allocation

Let $[X_1, X_2, \ldots, X_k]$ be a vector of k independent Gaussian random variables with variances $[\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2]$, and $[w_1, w_2, \ldots, w_k]$ a vector of positive weights that sum to one ($\sum_{i=1}^{k} w_i = 1$). Further assume without loss of generality that the X_i are in order of increasing variance:

$$\sigma_1^2 \leq \sigma_2^2 \leq \ldots \leq \sigma_k^2. \tag{A.1}$$

The distortion rate functions of the individual rv's are the k functions

$$D_i(R_i) = \sigma_i^2 \exp(-2R_i).$$

Now the *weighted rate allocation* problem is to find a vector $[R_1, R_2, \ldots, R_k]$ of non-negative rates that minimizes the weighted distortion

$$D = \sum_{i=1}^{k} w_i D_i(R_i) = \sum_{i=1}^{k} w_i \sigma_i^2 e^{-2R_i}, \tag{A.2}$$

subject to the rate constraint ($R > 0$)

$$\sum_{i=1}^{k} w_i R_i \leq R. \tag{A.3}$$

Temporarily ignoring the non-negativity constraints $R_i \geq 0$ we can set up a Lagrangian

$$J = \sum_{i=1}^{k} w_i \sigma_i^2 e^{-2R_i} + \lambda \left(\sum_{i=1}^{k} w_i R_i - R \right) \tag{A.4}$$

and set its partial derivatives to zero:

$$\frac{\partial}{\partial R_i} J = -2 w_i \sigma_i^2 e^{-2R_i} + \lambda w_i \stackrel{!}{=} 0. \tag{A.5}$$

Since the weights w_i are positive, the solutions of (A.5) do not depend on them:

$$R_i = \tfrac{1}{2} \ln \left(\frac{\sigma_i^2}{\lambda/2} \right). \tag{A.6}$$

Inserting these back into (A.2) yields the distortion $D = \sum_{i=1}^{k} w_i \sigma_i^2 \frac{\lambda/2}{\sigma_i^2} = \frac{\lambda}{2}$ and therefore (A.6) is equivalent to $R_i = \frac{1}{2} \ln(\frac{\sigma_i^2}{D})$. At this point we go back to the non-negativity constraints $R_i \geq 0$ and establish that the condition

$$D \leq D_{min} = \min\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2\} = \sigma_1^2 \tag{A.7}$$

suffices to satisfy all of them. Since D has now taken over the role of the Lagrange multiplier, we see that every value of $D \leq D_{min}$ solves the problem for a specific constraint (A.3), namely

$$R(D) = \sum_{i=1}^{k} \tfrac{1}{2} w_i \ln \left(\frac{\sigma_i^2}{D} \right) = \tfrac{1}{2} \ln \left(\prod_{i=1}^{k} \sigma_i^{2w_i} \right) - \tfrac{1}{2} \ln D. \tag{A.8}$$

Turning things around we also see that (A.7) corresponds to

$$R \geq R_{min} = \tfrac{1}{2} \ln \left(\prod_{i=1}^{k} \sigma_i^{2w_i} \right) - \tfrac{1}{2} \ln D_{min}. \tag{A.9}$$

The case when some R_i become zero can be solved via the Kuhn-Tucker conditions. Alternatively, we observe that (A.1) implies

$$R_1^2 \leq R_2^2 \leq \ldots \leq R_k^2.$$

Thus the sum (A.8) will become $R(D) = \sum_{i=i_0}^{k} \frac{1}{2} w_i \ln \left(\frac{\sigma_i^2}{D} \right)$ with the starting index $i_0 \in \{1, 2, \ldots, k\}$. A careful analysis reveals that for each rate constraint R and Lagrange multiplier λ there is an unique i_0, such that (A.3) is satisfied with equality and the positive rates R_i are given by (A.6). More details can be found e.g. in [28].

Bibliography

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

[2] Toby Berger. Optimum quantizers and permutation codes. *IEEE Trans. Inform. Theory*, IT-18:759–765, November 1972.

[3] Toby Berger and Jerry D. Gibson. Lossy source coding. *IEEE Trans. Inform. Theory*, IT-44:2693–2723, October 1998.

[4] Richard E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory*, IT-18:460–473, July 1972.

[5] James A. Bucklew and Neal C. Gallagher Jr. Quantization schemes for bivariate gaussian random variables. *IEEE Trans. Inform. Theory*, IT-25:537 – 543, September 1979.

[6] James A. Bucklew and Neal C. Gallagher Jr. Two-dimensional quantization of bivariate circularly symmetric densities. *IEEE Trans. Inform. Theory*, IT-25:667–671, November 1979.

[7] O. Cappé. A set of matlab/octave functions for the EM estimation of hidden markov models with gaussian state-conditional distributions, 1999. Available at `http://tsi.enst.fr/~cappe/mfiles/h2m.tar.gz`.

[8] Albert Cohen, Ingrid Daubechies, Onur Guleryuz, and Michael Orchard. On the importance of combining wavelet-based nonlinear approximation with coding strategies. preprint, 1997.

[9] Robert M. Corless, David J. Jeffrey, and Donald E. Knuth. A sequence of series for the Lambert W function. In *Proc. ISSAC '97*, pages 197–204, Maui, 1997. Also at `http://www.apmaths.uwo.ca/~rcorless/papers/LambertW/`.

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

[11] H.S.M Coxeter. *Regular Polytopes*. Dover Publications, 1973.

[12] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based statistical signal processing using hidden markov models. *IEEE Trans. Signal Proc.*, 46:886–902, April 1998.

[13] Aaron Deever and Sheila S. Hemami. What's your sign?: Efficient sign coding for embedded wavelet image coding. In *Proc. Data Compression Conference*, pages 273–282, Snowbird, Utah, March 2000.

[14] David L. Donoho, Martin Vetterli, R.A. DeVore, and Ingrid Daubechies. Data compression and harmonic analysis. *IEEE Trans. Inform. Theory*, IT-44:2435–2476, October 1998.

[15] Nariman Farvardin and James W. Modestino. Optimum quantizer performance for a class of non-gaussian memoryless sources. *IEEE Trans. Inform. Theory*, pages 485–497, May 1984.

[16] Stephen L. Fix. *Rate Distortion Functions for Continuous Alphabet Memoryless Sources*. PhD thesis, University of Michigan, September 1977.

[17] Stephen L. Fix. Rate-distortion functions for squared error distortion measures. In *Proc. 16th Annu. Allerton Conf. Communication, Control and Computers*, pages 704–711, Monticello, Illinois, 1978.

[18] Allan M. Gerrish and Peter M. Schultheiss. Information rates of non-gaussian processes. *IEEE Trans. Inform. Theory*, IT-10:265–271, October 1964.

[19] Allen Gersho. Asymptotically optimal block quantization. *IEEE Trans. Inform. Theory*, IT-25:373–380, July 1979.

[20] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

[21] Herbert Gish and John N. Pierce. Asymptotically efficient quantizing. *IEEE Trans. Inform. Theory*, IT-14:676–683, September 1968.

[22] Robert M. Gray. A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions. *IEEE Trans. Inform. Theory*, IT-19:480–489, July 1973.

[23] Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Trans. Inform. Theory*, IT-44:2325–2383, October 1998.

[24] Barry G. Haskell. The computation and bounding of rate-distortion functions. *IEEE Trans. Inform. Theory*, IT-15:525–531, September 1969.

[25] Are Hjørungnes and John M. Lervik. Jointly optimal classification and uniform threshold quantization in entropy constrained subband image coding. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, volume 4, pages 3109–3112, 1997.

[26] N. S. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice-Hall, Englewood Cliffs, NJ, 1984.

[27] John C. Kieffer. A survey of the theory of source coding with a fidelity criterion. *IEEE Trans. Inform. Theory*, IT-39:1473–1490, September 1993.

[28] John C. Kieffer. Lectures on source coding, 1997–2000. `http://www.ee.umn.edu/users/kieffer/ece5701.html`.

[29] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, IT-28:129–137, March 1982. First appeared as Bell Laboratories Technical Note in 1957.

[30] Daniel F. Lyons. *Fundamental Limits of Low-Rate Transform Codes*. PhD thesis, University of Michigan, August 1992. Available at `ftp://ftp.eecs.umich.edu/people/neuhoff/dlyons.thesis.ps`.

[31] Stéphane Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.*, 11:674–693, July 1989.

[32] Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[33] Stéphane Mallat and Frédéric Falzon. Analysis of low bit rate image transform coding. *IEEE Trans. Signal Proc.*, 46:1027–1042, April 1998.

[34] Michael W. Marcellin, Michael J. Gormish, Ali Bilgin, and Martin P. Bolick. An overview of JPEG-2000. In *Proc. Data Compression Conference*, pages 523–541, Snowbird, Utah, March 2000.

[35] J. Max. Quantizing for minimum distortion. *IRE Trans. Inform. Theory*, IT-6:7–12, March 1960.

[36] Peter Noll and Rainer Zelinski. Bounds on quantizer performance in the low bit-rate region. *IEEE Trans. Commun.*, COM-26:300–304, February 1978.

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

[38] Kenneth Rose. A mapping approach to rate-distortion computation and analysis. *IEEE Trans. Inform. Theory*, IT-42:1939–1952, November 1994.

[39] Hanan Rosenthal and Jacob Binia. On the epsilon entropy of mixed random variables. *IEEE Trans. Inform. Theory*, IT-34:1110–1114, September 1988.

[40] Martin C. Rost and Khalid Sayood. The root lattices as low bit rate vector quantizers. *IEEE Trans. Inform. Theory*, 1988.

[41] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition, 1976.

[42] A. Said and W.A. Pearlman. A new fast and efficient image codec on set partitioning in hierarchical trees. *IEEE Trans. on CSVT*, 6:243–250, June 1996.

[43] Khalid Sayood. *Introduction to Data Compression*. Morgan Kaufmann, San Francisco, CA, 2nd edition, 2000.

[44] Claude E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27:379–423, 623–656, July and October 1948. Also in *Claude Elwood Shannon: Collected Papers*, N.J.A Sloane and A.D. Wyner, Eds., IEEE Press 1993, pp. 5-83.

[45] Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE Conv. Rec.*, volume 7, pages 142–163, 1959. Also in *Claude Elwood Shannon: Collected Papers*, N.J.A Sloane and A.D. Wyner, Eds., IEEE Press 1993, pp. 325-350.

[46] Peter F. Swaszek and John B. Thomas. Multidimensional spherical coordinates quantization. *IEEE Trans. Inform. Theory*, IT-29:570–576, July 1983.

[47] Martin Vetterli and Jelena Kovacevic. *Wavelets and Subband Coding*. Prentice Hall, 1995.

[48] Claudio Weidmann and Martin Vetterli. Rate-distortion analysis of spike processes. In *Proc. Data Compression Conference*, pages 82–91, Snowbird, Utah, March 1999.

[49] Claudio Weidmann and Martin Vetterli. Rate distortion behavior of threshold-based nonlinear approximations. In *Proc. Data Compression Conference*, pages 333–342, Snowbird, Utah, March 2000.

[50] Jacob Ziv. On the best finite-state approximation of a stationary probability measure and applications. Presentation at the Third ETH-Technion Workshop on Information Theory held in Zurich, January 2000.

Curriculum Vitae

Claudio Weidmann
CH-7742 Poschiavo, Switzerland

10th October 1967 Born in Zürich, Switzerland.

1974-1987 Primary and secondary school in Zürich (1974-76) and Poschiavo (1976-82), then Lyceum Alpinum Zuoz (1982-87); obtained Matura Type C.

1987-1996 Software and hardware engineer at large with Ardesq Systems AG. Co-founder of follow-up company Arpage Systems AG (1992).

1988-1993 Study of Electrical Engineering at ETH Zürich; obtained diploma degree Dipl. El.-Ing. ETH.

1996-2000 Research assistant at the Audiovisual Communications Laboratory at EPF Lausanne; enrolled as PhD student.