TECHNIQUES CAUSALES DE CODAGE AVEC ET SANS PERTES POUR LES SIGNAUX VECTORIELS

- - -

CAUSAL LOSSY AND LOSSLESS CODING OF VECTORIAL SIGNALS

THÈSE NUMÉRO XXXX (2003)

Présentée au département **Signal et Images**,

ÉCOLE NATIONALE SUPÉRIEURE DE TÉLÉCOMMUNICATIONS

pour l'obtention du grade de DOCTEUR ÈS SCIENCES

par

David MARY

Composition du Jury:

J.-C. Pesquet, rapporteur

I. Tabus, rapporteur

T. Liebchen, examinateur

O. Michel, examinateur

G. Richard, examinateur

D. Slock, directeur de Thèse

Paris, ENST
Mars 2003

à ma Camille

Causal Lossy and Lossless Coding of Vectorial Signals

D. Mary

The aim of source coding, or compression, is to reliably represent some information by means of bits, with the natural concern for using a small number of bits. If the information can be exactly recovered from these bits, the code is called lossless; otherwise it is lossy. Both lossy and lossless coding are of interest in this work. Compression allows one to save bandwidth for data transmission over communications channels, or memory space for information storage.

The *information* considered in this thesis will be represented by vectorial signals, which compose a wide class of signals, among which scalar and multichannel signals. Multichannel signals may be obtained as soon as scalar signals are, in the context of various applications, gathered together. If this signals present some dependencies, such as audio signals for example, one should code them jointly in order to achieve a more efficient compression.

The initial idea of developping coding techniques for audio signals[1] motivated this choice of a vectorial representation. Though some applications will be presented for this kind of signals, the Gaussian assumption is often made since it allows one to derive closed form expressions, to compare, and possibly to prove the optimality of the considered coding schemes.

The first part of this thesis presents lossy coding techniques for vectorial signals.

In a transform coding framework firstly, we derive the optimal (linear) transform subject to the constraint of causality. This transform is shown to correspond to an LDU (Lower-Diagonal-Upper) factorization of the signal covariance matrix. This triangular transform is then compared to the Karhunen-Loève Transform (KLT), which is the optimal unitary transform for Gaussian signals, and which is therefore traditionally used as a benchmark. One criterion of merit used for this comparison is the coding gain, which corresponds to the ratio by which the distortion is decreased when using a particular transformation. Similarly as in DPCM (Difference Pulse Code Modulation), we show that practical causal coding schemes should be implemented in closed loop around the quantizers and, as in DPCM also, we show that at low rates a quantization noise feedback decreases the coding performance. For moderate to high rates however, we show that the optimal causal transform yields the same coding gain as its unitary counterpart. The optimal causal transform presents furthermore several advantages with respect to the KLT, such as lower implementation and design complexities, and perfect reconstruction property in the case of quantization of the transform coefficients.

In most of practical coding situations however, the data are nonstationary, which poses the problem of the adaptation of signal dependent transforms such as KLT or LDU. The main advantage of backward over forward adaptive coding schemes is to update the coding parameters with the data available at the decoder, avoiding thereby any excess bit rate. The coding performance of the two transformations are thus compared in this framework. This analysis allows one to quantitatively describe the influence of estimation and quantization noise as compared with the ideal case where the statistics of the signal are known.

Finally, the LDU transform is extended to (matricial) filtering in the last chapter of this first part. In this case, the optimal causal decorrelating scheme can be described by means of a prediction matrix, whose

[1] The first results of this work were obtained in the framework of the french RNRT project *COBASCA* :COdage en Bande élargie avec partage Adaptatif du débit entre Source et CAnal pour réseaux cellulaires de deuxième et troisième générations (UMTS).

entries are optimal prediction filters. The diagonal filters are scalar intrasignal prediction filters, and the off-diagonal predictors are Wiener filters performing the intersignal decorrelation. By considering vectors of infinite size, one can get frequential expressions for the coding gains. We show that this decorrelating scheme leads to the notion of "generalized" MIMO (Multiple Input Multiple Output) prediction, in which a certain degree of non causality may be allowed for the off-diagonal prediction filters. In the case of non causal intersignal filters, the optimal MIMO predictor is still lower triangular, and hence "causal", in a wider sense. The notion of causality may be generalized : the causality between channels becomes processing the channels in a certain order. Some signals may be coded using the coded/decoded versions of the "previous" signals. An interesting result is that if the quantization noise feedback is taken into account, the triangular predictor is the more efficient. Moreover, the coding gain is maximized if the signals are decorrelated by order of decreasing variance.

The second part of this thesis investigates lossless coding techniques, based on the previously considered causal approaches.

Recent work has shown that coding schemes using a lossless (integer-to-integer) implementation of the Karhunen-Loève Transform followed by scalar entropy coders are almost as efficient as vector entropy coders. We compare the integer-to-integer implementations of the KLT and LDU in this framework, which we refer to as "single-stage" lossless transform coding. We define the lossless coding gain for a transformation as the bitrate reduction operated by the corresponding lossless coding scheme. In a first step, we show that the maximal achievable coding gain corresponds to the average mutual information shared by the components of the vector. In a second step, we analyze the effects of the integer-to-integer constraint on the coding gains. A third step analyzes the effects of estimation noise uppon the coding gains : in this case, the transforms are based on an estimate of the covariance matrix of the quantized signals. We find that for stationary Gaussian signals, the coding gains are close to their maxima after a few tens of decoded vectors. Moreover, because of its triangular structure, the LDU based approach is shown to yield the highest coding gain.

Orthogonal transforms are then compared with the causal transform in "multi-stage" lossless transform coders. For internet browsing applications, or in the case of varying transmission bandwidth, this kind of schemes allows one to deliver in a first step a low resolution (lossy) version of the signal, and to transmit separately the error signal. In a two-stage lossless coder, each vector is transformed, quantized, and an error signal is generated by substraction to the original signal. For orthogonal transforms, the cost of the multiresolution approach is a bitrate penalty of 0.25 bit per sample. This excess bitrate is due to a "gaussianization effect" of the transforms. We show that the causal approach allows one, in this framework, to code the data (almost) without causing any excess bitrate as compared with a single-stage coder. Also, the approach based on the causal transform allows one to easily switch between a single- or a multi-stage compressor. Moreover, the proposed approach allows one to easily fix the distortion and rate for both the low resolution and the error signal in each channel. Any of the channels may, as a particular case, be chosen to be directly losslessly coded.

Finally, we apply our results about optimal coding of vectorial signals to the single- and multi-stage lossless structures described so far. In a first step, prediction matrices of the generalized MIMO prediction framework are used in single-stage coders. The corresponding compression performances are compared to the optimal compression performances, as achievable by any lossless coding technique. The particular

cases of the classical and the triangular MIMO predictors are investigated, and shown to present equivalent performances. In a second step, we investigate the performances of two-stage structures where (A)DPCM loops are introduced. The quantizers of these loops allow one to choose the respective bitrates for both the error and the low resolution signals. For these two-stages structures, the overall bitrate delivered by the multiresolution structure is compared to that of the corresponding "one-shot" approach. These two-stage structures are shown to be slightly suboptimal because of the noise feedback created in the (A)DPCM loops. Finally, we show that the two-stage structure can easily be extended to a larger number of stages. In that case, a simple method is proposed so that the delivered bitrates approach some predetermined target rates.

Techniques Causales de Codage avec et sans Pertes pour les Signaux Vectoriels

D. Mary

La nécéssité de “comprimer” les signaux numériques trouve son origine dans les moyens limités dont disposent les communications numériques : la compression permet d’économiser la bande passante des canaux sans-fil ou internet; elle permet aussi d’économiser l’espace mémoire en ce qui concerne leur stockage. D’une façon générale, le codage de source consiste à mettre au point des techniques permettant, suivant l’application visée, de déterminer le meilleur compromis entre la qualité avec laquelle les informations seront représentées, et la ressource, ou le débit, qui sera nécessaire pour décrire la représentation choisie. Selon que l’information initiale peut être partiellement, ou parfaitement reproduite après l’opération de codage, on parle de codage avec, ou sans pertes. Cette thèse présente diverses techniques, et l’évaluation de leur efficacité, pour ces deux types de codage.

L’ *information* considérée dans cette thèse sera représentée par des signaux vectoriels, qui forment une large classe de signaux, incluant par exemple les signaux scalaires ou les signaux multicanaux. Ces derniers peuvent être construits dès que plusieurs signaux scalaires sont, pour des applications diverses, regroupés. Dès lors que les signaux scalaires individuels présentent des dépendances, comme certains signaux audio par exemple, il y a un intérêt à les traiter conjointement, en vue d’une compression plus efficace.

L’idée initiale de développer des techniques adaptées aux signaux audio[2] a motivé ce choix d’une représentation vectorielle. Bien que quelques applications soient présentées pour ce type de signaux, l’hypothèse de signaux gaussiens est souvent retenue, car elle permet d’obtenir des résultats analytiques et donc de comparer et de prouver, le cas échéant, l’optimalité des schémas de codage considérés.

La première partie de cette thèse présente des techniques de codage avec pertes pour les signaux vectoriels.

Dans le cadre du codage par transformée tout d’abord, nous nous intéressons au codage de signaux vectoriels par une transformation décorrélatrice causale de type DPCM (Differential Pulse Code Modulation, technique utilisée pour les signaux scalaires, supprimant les redondances par prédiction linéaire). Nous montrons que la transformation causale optimale correspond à une factorisation triangulaire LDU (Lower-Diagonal-Upper) de la matrice d’autocorrélation du vecteur de signal à coder. Cette approche est comparée à sa contrepartie unitaire, la transformation de Karhunen-Loève (KLT), bien connue parce qu’étant optimale pour les sources gaussiennes, elle sert traditionnellement de référence. Plusieurs aspects sont abordés dans cette comparaison, comme le gain de codage apporté par la transformation (qui correspond au facteur par lequel la distortion est réduite, pour un même débit, grâce à la transformation), les effets intervenants lorsque le schéma de codage est implémenté en boucle fermée (c’est à dire lorsque la transformation utilise des données précédemment quantifiées, ce qui introduit dans le schéma de codage un retour de bruit), ou la complexité algorithmique. Nous proposons une analyse des perturbations liées au retour de bruit, qui montre que quand celui-ci devient négligeable, les performances sont identiques à celles obtenues dans le

[2] Les premiers résultats de ce travail ont été obtenus dans le cadre du projet RNRT *COBASCA* : COdage en Bande élargie avec partage Adaptatif du débit entre Source et CAnal pour Réseaux cellulaires de deuxième et troisième générations (UMTS).

cas unitaire, bien que la complexité de la LDU soit notablement moindre. Dans la plupart des cas pratiques cependant, les données réelles sont non stationnaires, ce qui pose un problème d'adaptation pour des transformations dépendant du signal telles que la KLT ou la LDU. Nous étudions donc les performances de schémas de codage pour lesquels ces transformations sont adaptées sur la base de données quantifiées, ce qui évite un surcroît de débit qui correspondrait à transmettre au décodeur une description de ces transformations. Dans ce contexte, nous analysons les effets de perturbation liés au bruit de quantification et au bruit d'estimation qui se posent par rapport au cas idéal. Cette analyse permet d'évaluer quantitativement, en fonction d'un débit moyen imposé et du nombre de données précédemment décodées, l'écart entre la performance réelle des deux systèmes et leur performance idéale, où les statistiques des signaux à compresser sont connues.
Dans la fin de cette première partie, l'approche matricielle causale de type LDU est généralisée au cas où les coefficients de la matrice de transformation triangulaire sont des filtres prédicteurs (prédiction MIMO -Multi Input Multi Output- triangulaire). Cette généralisation débouche sur la prédiction MIMO dite “généralisée”, pour indiquer que la prédiction MIMO classique et la prédiction MIMO triangulaire constituent deux cas particuliers, parmi une infinité, d'une même approche totalement décorrélatrice, et “causale” dans un sens plus large. Nous montrons que si les effets de retour de bruit de quantification sont pris en compte, la prédiction MIMO triangulaire est, parmi toutes ces approches, celle qui maximise le gain de codage. Dans ce cas, décorréler les signaux par ordre de variance décroissante est optimal. Une application de ces résultats est proposée dans le cadre du codage de la parole large bande ([0-7kHz]).

La deuxième partie de cette thèse développe des techniques de codage sans pertes basées sur les approches causales considérées précédemment.
Une première étape consiste à comparer les performances de la LDU à celles de la KLT dans le cas où elles sont implémentées de façon à être sans pertes (transformations “d'entiers à entiers”). Le gain correspond alors à la réduction de débit opérée par la transformation, tout en garantissant une représentation exacte de la source. Nous montrons d'abord que le gain maximal qui peut être apporté par de telles transformations correspond à la moyenne des informations mutuelles partagées par les différentes variables qui composent le processus vectoriel. Nous analysons ensuite les gains apportés par la KLT et la LDU dans ce cadre, et décrivons notamment les effets dûs à la contrainte “entiers à entiers” en terme de débit supplémentaire par rapport au gain idéal. Le bruit d'estimation pour un schéma adaptatif est aussi traité. L'approche causale, grâce à sa nature triangulaire, s'avère présenter dans ce cadre des performances légèrement supérieures à l'approche unitaire.
Nous étudions ensuite des schémas de codage sans pertes qui permettent de délivrer, dans un premier temps, une version basse résolution du signal d'intérêt, et de transmettre le signal complémentaire par la suite. Ce genre de schéma est utile pour des applications de navigation rapide sur internet, ou de transmission à bande passante variable. La transformation causale est comparée dans ce cadre aux transformations orthogonales. Nous considérons une version légèrement modifiée d'un schéma à deux niveaux de résolution simple (utilisé par exemple dans le contexte du codage audio sans pertes), dans lequel chaque vecteur est d'abord transformé, quantifié, puis transmis comme version basse résolution du signal. Un signal d'erreur est ensuite généré par soustraction au signal original, et transmis comme complément. L'extension de ce schéma à plusieurs niveaux de résolution est obtenue en introduisant des quantificateurs de type APCM dans le schéma sans pertes. On montre que les transformations orthogonales classiques sont sous-optimales pour

de telles approches multirésolution par rapport à leur alternative causale. La transformation causale présente d'autres avantages par rapport à des transformations telles que la KLT ou la DCT, comme la possibilité de passer instantanément d'un schéma de codage sans pertes monorésolution, à des schémas multirésolution, de pouvoir choisir des niveaux de résolution différents pour chacun des canaux et, notamment, de pouvoir coder sans pertes un ou plusieurs canaux particuliers uniquement.

Finalement, des schémas de codage sans pertes multirésolutions sont présentés, qui se basent sur la prédiction MIMO considérée dans la première partie. Nous montrons que l'approche multirésolution est légèrement sous-optimale en terme de débit total par rapport à une approche de compression globale à cause du retour de bruit dans les boucles de type ADPCM. On propose aussi une méthode pour que les débits générés par chacune des résolutions correspondent à des débits cibles prédeterminés.

Contents

List of Figures

Acronyms and Abreviations

ascii	american standard code for information interchange
JPEG	Joint Photographic Expert Group
2D	two dimensions
3GPP	3rd Generation Partnership Project
AR(n)	Autoregressive Process of order n
COBASCA	COdage en Bande élargie avec partage Adaptatif du débit entre Source et CAnal pour réseaux cellulaires de deuxième et troisième générations (UMTS), RNRT project, http://www.telecom.gouv.fr/rnrt/pcobasca.htm
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
(A)DPCM	(Adaptive) Difference Pulse Code Modulation
DST	Discrete Sine Transform
DWT	Discrete Wavelet Transform
ECUQ	Entropy Coded Uniform Quantizer (or Quantization)
ETSI	European Telecommunications Standards Institute
I2I	Integer-to-Integer
FIR	Finite Impulse Response
FRQ	Fixed-Rate Quantization
GSM	Global System for Mobile communications
i.i.d.	independent and identically distributed
KLT	Karhunen-Lòeve Transform
LDU	Lower-Diagonal-Upper, causal transform

LMS	Least Mean Squares (adaptive filtering)
MIMO	Multiple Inputs Multiple Outputs
MISO	Multiple Inputs Single Output
(L)MMSE	(Linear) Minimum Mean Square Error
MPEG	Moving Picture Expert Group
MSE	Mean Squared Error
PCM	Pulse Code Modulation
p.d.f.	probability density function
psd	power spectral density
QMF	Quadrature Mirror Filters
CQF	Conjugate Quadrature Filters
RLS	Recursive Least-Squares (adaptive filtering)
RLSF	Recursive Least-Squares-Fitting (channel estimation technique)
RNRT	Réseau National de Recherche en Télécommunications
SBC	Subband Coding
SNR	Signal-to-Noise Ratio
SISO	Single Input Single Output
TC	Transform Coding
UMTS	Universal Mobile Telecommunication System
(V)HR	(Very) High Resolution
VRQ	Variable-Rate Quantization
w.l.g.	without loss of generality
w.r.t.	with respect to
WSS	Wide Sense Stationary

List of Symbols

x	Scalar random variable, process, source or signal
x_k	Particular outcome, or sample of x (*e.g.* sample of x at time instant k)
$\underline{x}$	Vectorial r.v., process, source or signal
$\underline{x}_k$	Particular outcome, or sample of $\underline{x}$ (e.g. sample of $\underline{x}$ at time instant k)
x_i	ith component of $\underline{x}$
$x_{i,k}$	Particular outcome, or sample of the ith component of $\underline{x}$
$\underline{x}_{i:j,k:K}$	Set of samples of the components $x_i, x_{i+1}\cdots x_{j-1}, x_j$ at instants $k, k+1\cdots K-1, K$
$\underline{x}^q$	Vector whose components are quantized
$\widetilde{\underline{x}}$	Reconstruction error vector
$\underline{X}_k$	Supervector $[\underline{x}_1\underline{x}_2\cdots\underline{x}_k]^T$
R or $R_{\underline{x}\underline{x}}$	Correlation matrix E $\underline{x}_k\underline{x}_k^T$
R_{ij}	Element of the matrix R in row i and column j
$\mathcal{P}$	Permutation matrix
diag$\{R\}$	Diagonal matrix with same diagonal as R
$\overline{\text{diag}}\{\underline{x}\}$	Diagonal matrix with $R_{ii} = x_i$
$\delta_{i,j}$	Kronecker symbol (equals 1 if $i = j$ and 0 otherwise)
I_N or I	$N \times N$ Identity matrix
j	$\sqrt{-1}$
$(\cdot)^*$	Complex conjugate operator
$(\cdot)^T$	Transpose operator
$(\cdot)^H$	Hermitian operator
$*$	Convolution operator
tr	Trace operator
$\det$	Determinant
vec	Vec operator
o	Composition operator
$\|\|A\|\|$	Euclidean norm of vectors and matrices, $(\,\mathrm{tr}\,\{A^H A\})^{\frac{1}{2}}$
$\|\|A\|\|_\infty$	Infinity norm, $\max\limits_{i,j} \|A_{ij}\|$
$\triangleright(.)\ (\mathrm{resp.}\ \triangleleft\,(.))$	Lower (resp. upper) triangular matrix made with the lower (resp. upper) triangular part of (.)
E	Mathematical expectation
$\otimes$	Kronecker product

Chapter 1

Introduction

Consider a system composed of two parts, a first part which possesses some information, and a second part which does not. If the first part is concerned in reliably transmitting information to the second one, then this can be called a *communication system.* This formulation is both abstract and general. Communication systems are ubiquitous. Providing a powerful mathematical framework to analyze communication systems remained for a long time a complex and unsolved problem. It was the seminal work of Shannon [2], who introduced a precise and flexible enough abstraction of a communication system, which launched this mathematical discipline, Information Theory.

In the last decades, Information Theory has provided many results about two original and fundamental problems. The problem of building from an original *message* a *signal*, which will efficiently represent the information of interest, once, more fundamentally, "efficiently" has been precisely defined, is called *the source coding* problem. The problem of characterizing and understanding the way this signal will be corrupted during the transmission, and proposing further operational systems which will effectively protect the information of interest, is referred to as the *channel coding problem*. However, beyond communications, many scientific fields were impacted by Information Theory, including Probability, Statistics, Computation Theory or Economics [3]. In the particular case of physics[1], some authors consider that information, as defined by Shannon, may be a fundamental concept [2], even more so than energy [5].

In the framework considered in this work, the "information of interest" may be any mathematical signal describing physical quantities, images, speech, or music signals. In practice, transmission of these information was analog (continuous-time and continuous-amplitude signals) up to the second half of the last century. Since the introduction of Pulse Code Modulation (PCM) however, communication is almost al-

[1] As explicitely stated in [2], the difference between Boltzmann's and Shannon's entropy merely amounts to a choice of a unit of measure.

[2] In [4], classical and quantum particle statistics are rederived using information theoretic arguments.

ways, and increasingly, digital. The PCM system was historically patented in 1938 [6], used for military communications systems in 1945 [7] and published in 1947 [8, 9]. Because PCM systems perform on analogic signals a double discretization, in time (sampling) and amplitude (scalar quantization) they are also referred to as A/D (Analog to Digital) converters. The main advantage of digital communication systems is that, by introducing some loss (due to double discretization) in a controlled fashion, further loss can be prevented during the transmission. The information is then transmitted by means of information elements, the bits, resulting in a certain bitrate. Very generally, Source Coding deals with representing some information by means of bits. If the original information can be exactly recovered from these bits, the coding is called lossless, otherwise it is lossy. The branch of Information Theory which is dedicated to the problem of characterizing the minimum rate required to represent a source up to a certain resolution level is called Rate-Distortion Theory. Both lossy and lossless coding techniques will be of interest in this work.
The aim of the following introduction is to set the mathematical framework of this dissertation, and to recall some historical results. The particular topics of interest, and the main purposes of this work should be underlined along the mathematical setting, and will be more precisely exposed at the end of the chapter.

1.1 About Shannon's Mathematical Theory of Communication

The mathematical abstraction of a communication system as proposed by Shannon in [2] is represented by Figure 1.1. In this abstraction, an *information source* produces a *message*, or a sequence of messages, which may generally be continuous- or discrete- time and amplitude. The *transmitter* operates on the message to produce a signal suitable for transmission over the channel. This *channel* represents the physical medium used to transmit the signal (wires, RF spectrum, fiber optical [10]...). The *receiver* attempts to recreate the message from the received signal, and delivers this message to the *destination*, which is the person or thing to whom the message is intended.

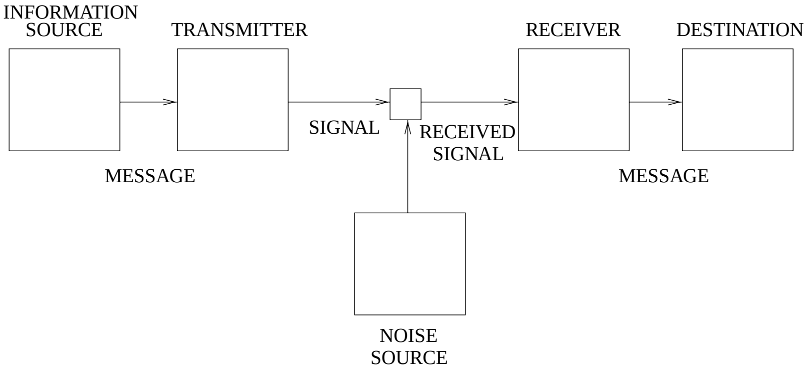


Figure 1.1: Shannon's schematic diagram of a general communication system.

One achievment of Shannon's description is to represent the various elements involved in this description by mathematical entities, idealized from their physical counterparts. We should now present the definitions and notations of important quantities, and briefly recall some important results, which will be relevant to this dissertation.

1.2 Definitions and Important Results

This short presentation aims only of introducing the results of the next section. Several properties and interpretations regarding entropy, relative entropy and mutual information can for example be found in [3].

1.2.1 Information

Let $i(n)$ be a stationary random process described by series of discrete independent and identically distributed (i.i.d.) random variables (r.v.s) with alphabet $\mathcal{I}$. This source is called a *discrete memoryless* source. We denote by

$$p_i(i_k) = \mathrm{Pr}\{i(n) = i_k\}, \tag{1.1}$$

the distribution of the several probabilities. Each outcome $\{i(n) = i_k\}$ contains an information

$$I(i_k) = -\log_2 p_i(i_k). \tag{1.2}$$

Since we choose a logarithm of base 2 for the definition, I is expressed in bits per symbol. The lower the probability, the higher the information: in some sense, being informed is being surprised.

1.2.2 Entropy

The discrete entropy of $i(n)$ is defined as the mathematical expectation of the r.v. $I(i_k)$,

$$H(i) = \mathrm{E}\, I(i_k) = -\sum_{i_k \in \mathcal{I}} p_i(i_k) \log_2 p_i(i_k). \tag{1.3}$$

The entropy may be interpreted as the average quantity of information delivered by an outcome of $i(n)$, and measures the amount of uncertainty associated with the source. $H(i)$ correponds also to the minimal number of bits required to exactly describe the discrete memoryless source $i(n)$. The entropy of a discrete source can be shown to be positive or null, and upper bounded by $\log_2 N_{\mathcal{I}}$, where $N_{\mathcal{I}}$ is the number of elements of $\mathcal{I}$. Hence, entropy is maximal (and equals $\log_2 N_{\mathcal{I}}$) if all the symbols are equiprobable, see e.g. [11].

For a continuous real-valued r.v. *i* with p.d.f. p_i, the *differential* entropy $h(i)$ of i is defined as

$$h(i) = -\int_{-\infty}^{+\infty} p_i(\mathrm{i}) \log_2 p_i(\mathrm{i}) d\mathrm{i}. \tag{1.4}$$

Historic references about the concept and the origin of entropy, the relationship between differential and discrete entropy, and particular applications to coding of audio signals can be found in [12].

1.2.3 Entropy Rate

The previous definition can be generalized to the case of N discrete r.v.s. Let us consider the vector $\underline{i}(n) = [i(n)\ i(n+1) \cdots i(n+N-1)]^T$, and define by $p_{\underline{i}}(i_{1k}, \cdots, i_{Nk})$ the joint probability

$$p_{\underline{i}}(i_{1k}, \cdots, i_{Nk}) = \mathrm{Pr}\{i(n) = i_{1k}, \cdots, i(n+N-1) = i_{Nk}\}. \tag{1.5}$$

The entropy of this vector is defined as

$$H(\underline{i}) = -\sum_{i_{1k}\cdots i_{Nk}} p_{\underline{i}}(i_{1k},\cdots,i_{Nk})\log_2 p_{\underline{i}}(i_{1k}\cdots i_{Nk}). \tag{1.6}$$

The *entropy rate* $H_\infty(i)$ of the process $i(n)$ is then the limit

$$H_\infty(i) = \lim_{N\to\infty}\frac{1}{N}H(\underline{i}). \tag{1.7}$$

One can show that it is possible to build an uniquely decodable code to a source as long as the number of bits per symbol is at least as high as the entropy rate of the source.
Also, (1.4) and (1.7) may be generalized to the case where $\underline{i}(n) = [i(n)\ i(n+1)\cdots i(n+N-1)]^T$ is composed of N continuous r.v.s. For $\underline{i}$ having joint p.d.f. $p_{\underline{i}}$, the differential entropy of $\underline{i}$ is defined as

$$h(\underline{i}) = -\int p_{\underline{i}}(\mathrm{i})\log_2 p_{\underline{i}}(\mathrm{i})d\mathrm{i}, \tag{1.8}$$

and the corresponding *differential entropy rate* of the continuous source $i(n)$ as

$$h_\infty(i) = \lim_{N\to\infty}\frac{1}{N}h(\underline{i}). \tag{1.9}$$

1.2.4 Mutual Information

Let i and j be two discrete r.v.s with respective alphabets $\mathfrak{I}$ and $\mathfrak{J}$. The mutual information $I(i;j)$ is the average reduction in uncertainty of an event $\{i(n) = i_k\}$ due to the knowledge of an event $\{j(n) = j_k\}$, and is defined by

$$I(i;j) = \sum_{i_k\in\mathfrak{I}}\sum_{j_k\in\mathfrak{J}} p(i_k,j_k)\log_2\frac{p_{ij}(i_k,j_k)}{p_i(i_k)p_j(j_k)}. \tag{1.10}$$

The mutual information is nonnegative and corresponds to the relative entropy between the joint probability and the product of the marginal probabilities.

1.2.5 Capacity

Suppose now that the values of process $i(n)$ correspond to the input symbols of a channel, and that the output symbols $\{j_k\}\in\mathfrak{J}$ depend only on the input symbol i_k at the same instant. For this so-called *discrete and memoryless* channel, the capacity C is defined by

$$C = \max_{p_i} I(i;j). \tag{1.11}$$

The fundamental theorem for a discrete channel with noise ([2],Th. 11) states that communication with arbitrarily low error probability is possible if, and only if

$$H(x)\le C. \tag{1.12}$$

This theorem is sometimes referrred to as the “separation theorem” for stationary memoryless sources and channel. Extensions to other sources and channels are reviewed in [13]. This theorem and the results of [13] suggest that one could design practical methods to compress a source without any knowledge of the channel. Similarily, the design of a system aiming of communicating over a particular channel could be designed without regard for the significance of each particular bit. For this reason, source and channel coding have grown into separate fields with rather separate communities. Because of channel capacity variations however, joint source-channel coding may improve the overall coding performance of real systems. Thus, the optimality of designing separately source and channel coders is somehow idealistic, but leads however to easier designs.

As stated as early as 1984 in [14], digital source coding is by no means a new topic. One may however hope that there is still room for improvement and innovation. A primary purpose of this work is indeed to show that this is the case.

1.3 Source Coding : General Presentation

A source code is composed of two mappings: an *encoder* and a *decoder*, see Figure 1.2.

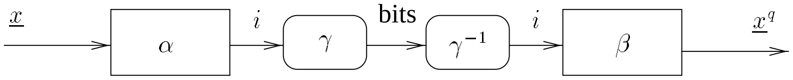


Figure 1.2: General representation of a source code.

The encoder maps any vector $\underline{x}_k \in \mathcal{R}^N$ to a finite string of bits, and the decoder maps any of these strings of bits to an approximation $\underline{x}_k^q \in \mathcal{R}^N$. The encoder can always be factored as $\gamma \circ \alpha$, where α is a mapping from $\mathcal{R}^N$ to some discrete set $\mathcal{I}$, and γ is an invertible mapping from $\mathcal{I}$ to strings of bits. Operations α and β are referred to as lossy encoder and decoder, and define a *quantizer*. The operation γ is called a *lossless*, or *entropy code*.

The quality of a source code is assessed by measuring the approximation accuracy of $\underline{x}^q$ with respect to (w.r.t.) $\underline{x}$, and the length of the description. The measure for the description length will be the expected number of bits delivered by the encoder divided by the vector length N. This is called *the rate* in bits per scalar sample. The measure of approximation accuracy will be the expected squared Euclidian norm divided by the vector length

$$d(\underline{x}, \underline{x}^q) = \frac{1}{N}\,\mathrm{E}\,||\underline{x}_k - \underline{x}_k^q||^2 = \frac{1}{N}\,\mathrm{E}\,||\underline{\tilde{x}}_k||^2, \tag{1.13}$$

where $\mathrm{E}\,||\underline{\tilde{x}}_k||^2$ denotes the variances of the reconstruction error. The mean squared error (MSE) distortion as defined in (1.13) is very conventional and usually leads to the easiest mathematical results. Source coding theory has however been developped for quite general distortion measures [15].

Concerning on the one hand the lossy component of the source code, also called *quantization stage*, each $\underline{x}_k \in \mathcal{R}^N$ is mapped from a source alphabet to a reproduction codebook $\mathcal{C} = \{\underline{x}_i^q\}_{i \in \mathcal{I}} \subset \mathcal{R}^N$, where $\mathcal{I}$ is an index set. Quantization operation Q is then realized by cascading the operation α and β. The lossy encoder α: $\mathcal{R} \rightarrow \mathcal{I}$ is specified by a partition of $\mathcal{R}^N$ into partition cells $S_i = \{\underline{x}_k \in \mathcal{R}^N | \alpha(\underline{x}_k) = i\}, i \in \mathcal{I}$. The reproduction decoder β: $\mathcal{I} \rightarrow \mathcal{R}^N$ is specified by the codebook $\mathcal{C}$. If $N = 1$, the quantizer is called *scalar*, and otherwise *vector quantizer*. Most popular lossy coding techniques include (possibly predictive) scalar and vector quantization, transform or subband coding, and combinations thereof. Those of them which are involved in the coding structures of interest in this thesis will be briefly described in section 1.5. Extensive literature about each of them exist however; for a comprehensive overview, see the excellent tutorials [16] and [17].

On the other hand, the aim of the entropy coder is to assign a unique binary string called a codeword to each $i \in \mathcal{I}$. A trivial assignment consists of transmitting codewords which correspond to the binary representation of each index. Since the codewords have equal lengths, this procedure is called *fixed-rate* coding. The codeword assignment may however be done in such a way that the average bitrate is lower than in fixed-rate coding. The basic idea is to assign shorter codewords to the indexes whose cells are more frequently used by the quantization process, and longer codewords for indexes which are less likely. Indeed, lossless compression is achievable in this case only if the probabilities of selection of the quantization cells are different. This coding scheme is then referred to as *variable rate* coding. Though lossless coders may also exist as standalone coders, they are always required as parts of lossy coding schemes. Lossless coding techniques are reviewed in some more details in the next section.

1.4 Lossless coding

1.4.1 Introduction

Lossless coding is also called *data compaction, noiseless, invertible,* or *entropy* coding. As discussed above, lossless compression can be achieved for discrete sources emitting symbols in a finite alphabet by taking advantage of the non equal probabilities of occurence of the symbols. The cost is firstly some encoding delay allowing one to reliably estimate these probabilities. Secondly, if the average bitrate may be decreased by using variable rate coding, the instantaneous (or on a short period of time) bitrate may be arbitrarily high, which may cause buffer overflows. This means that applying lossless codes may result in data expansion instead of compaction in the short run. Finally, variable rate coding may suffer from error propagation if some bits are received by the decoder in error. The main advantage of lossless over lossy coding is indeed to guarantee, assuming a noiseless channel, that the data will be exactly recovered. In many applications, such as computer programming, bank statements, some medical applications..., nonperfect information recovering is not acceptable. In some more particular applications, such as audio archiving and mixing, lossless compression may also be desired. This kind of techniques will be investigated in the second part of this thesis.

1.4.2 Entropy Codes

We now precise some properties and definitions about entropy coding. Let us consider a discrete random variable i with alphabet $\mathcal{I}$. The entropy coder assigns a unique binary string called a codeword to each $i \in \mathcal{I}$ (Figure 1.2). Since the codewords are unique, entropy codes are always invertible. A code is called *uniquely decodable* if the output sequence $\gamma(i_1), \gamma(i_2), \cdots, \gamma(i_k)$ corresponding to the input sequence $i_1, \cdots, i_k$ is one-to-one. Uniquely decodable codes can be applied to message sequences without adding any “punctuation” sign to show where codewords begin and end. If no codeword is the prefix of any other codeword, the code is called a *prefix* code. Prefix codes are guaranteed to be uniquely decodable.

The expected code length corresponding to the entropy code γ is then

$$L(\gamma) = \mathrm{E}\,[l(\gamma(i))] = \sum_{i_k \in \mathcal{I}} p_i(i_k) l(\gamma(i_k)), \tag{1.14}$$

where $p_i(i_k)$is the probability of the symbol i_k and $l(\gamma(i_k))$ is the length of the corresponding codeword. The entropy code is *optimal* if it is a prefix code that minimizes $L(\gamma)$. Huffman codes [18] are examples of optimal entropy codes. The performance of an optimal code is bounded by

$$H(i) \leq L(\gamma) < H(i) + 1, \tag{1.15}$$

where H is the discrete entropy defined in (1.3). More useful upper bounds may be found, e.g. [11, 19, 20]. Concerning the lower bound, analytical formulas which would describe the rate given by Huffman codes as a function of the probabilities are unknown [16]. Approximating this rate by the entropy gives however a useful though underestimated idea of the actual achievable rates.

Among most famous examples of lossless codes are the Morse code of 1837 (where the binary representation is replaced by dots and dashes, and the codewords' lengths are inversely proportional to the letters relative frequencies; this code requires fewer bits than fixed-rate ascii), Huffman code (1952, used in Unix *compact* utility), run-length codes (popularized by Golomb in the early 1960's, and used in the JPEG standard), Golomb codes which are type of Huffman codes, Lempel-Ziv(-Welch) codes ($1977-78$, used in Unix *compress* utility), arithmetic codes, Rice code (Huffman code for Laplacian probability density functions, used in many state-of-the-art lossless audio coders [21, 22, 23, 24]).

1.5 Lossy Coding

The main results of Lossy Coding or Quantization come historically from two complementary approaches: the information theoretic approach of Shannon, also called rate-distortion theory or *source coding with a fidelity criterion* [2, 25], and the *high resolution*, or *high rate* or *asymptotic* theory, whose origin can be found in [26, 27, 28].

1.5.1 Rate-Distortion Function

In many practical cases, noiseless coding of discrete sources is not possible. One wishes to describe the performance of a system which allows one to compress the source by accepting some distortion. Considering a distortion measure $d(\underline{x}, \underline{x}^q)$, the *rate-distortion* function [15] describes the lowest rate required to represent a continuous source $\underline{x}$ (taking values in the N-dimensional Euclidean space $\mathbb{R}^N$) with distortion no greater than some maximum distortion D

$$r_{\underline{x}}(D) = \inf_{d(\underline{x}, \underline{x}^q) \leq D} \frac{1}{N} I(\underline{x}; \underline{x}^q), \tag{1.16}$$

where the infimum of the normalized mutual information $\frac{1}{N} I(\underline{x}; \underline{x}^q)$ is taken over all joint distributions of $\underline{x}$ and $\underline{x}^q$ such that $d(\underline{x}, \underline{x}^q) \leq D$. Alternatively, one can define a *distortion-rate function* which is the least distortion with rate $r_{\underline{x}}$ or less. For $\underline{x}$ having joint p.d.f. $p_{\underline{x}}$ and finite differential entropy $h(\underline{x})$ the *Shannon lower bound* states that for an MSE distortion measure

$$r_{\underline{x}}(D) \geq \frac{1}{N} h(\underline{x}) - \frac{1}{2} \log_2(2\pi e D). \tag{1.17}$$

One important feature of the Shannon lower bound is that it easily generalizes to stationary sources. Let x be a real stationary source, and let $\underline{x}$ denote the vector of the first N samples of x. The rate-distortion function of x is defined by

$$r_{x,\infty}(D) = \lim_{N\to\infty} r_{\underline{x}}(D) = \lim_{N\to\infty} \frac{1}{N}h(\underline{x}) - \frac{1}{2}\log_2(2\pi e D), \tag{1.18}$$

and the limit is known to always exist [15]. Assuming a differential entropy rate $h_\infty(x) = \lim_{N\to\infty} \frac{1}{N}h(\underline{x})$, the *generalized Shannon lower bound* is

$$r_{x,\infty}(D) \geq h_\infty(x) - \frac{1}{2}\log_2(2\pi e D). \tag{1.19}$$

The performance of realizable quantizers as designed by high resolution quantization theory may then be compared to these information theoretic results [3]. The quality of a quantizer, as defined in section 1.3, is determined by its distortion and rate. We will limit this review by considering MSE distortion (1.13). As for the rate corresponding to a particular distortion, it can be measured in a few ways. Associating a particular entropy code γ to the quantizer gives a *variable rate quantizer* (α, β, γ), whose rate is given by eq. (1.14), possibly divided by the length N in the vector case. If no particular code is specified, or if binary representation is used to build the codewords, the quantizer is called *fixed-rate quantizer*, whose rate is hence $\log_2 N_{\mathfrak{I}}$. The ideal case where the rate is measured by the entropy $H(i)$ yields an *entropy-constrained* quantizer. The optimal performance of variable-rate quantization is at least as good as that of fixed-rate quantization, and entropy-constrained quantization is better yet.

1.5.2 High Resolution Scalar Quantization

For most sources, it is impossible to analytically express the performance of optimal quantizers. The approximations obtained when it is assumed that the quantization is very fine are however reasonnably accurate even at low to moderate rates [30, 31]. See [16, 17] and references therein for a comprehensive overview of the main historical contributions to high rate quantization.

Let p_x denote the p.d.f. of a continuous scalar r.v. x. High resolution analysis is based on approximating p_x on the interval S_i by its value at the midpoint. Assuming p_x is smooth, this approximation is accurate when all S_i are short [4]. Optimizing a scalar quantizer turns into finding the optimal lengths for the cells, depending on the p.d.f.. For large rate, the performance of optimal fixed-rate quantization (FRQ) is approximately

$$\mathrm{E}\,\widetilde{x}_k^2 \approx \frac{1}{12}\left(\int_{-\infty}^{+\infty} p_x^{\frac{1}{3}}(x)dx\right)^3 2^{-2r}, \tag{1.20}$$

which is now called the Panter and Dite formula [28]. Optimal conditions for p.d.f. optimized quantizers require that the quantization follows a nearest reconstruction level rule, and that these levels are the conditional expectation of the source value given that it lies in the specified cell (also called *centroid of the p.d.f.* in the interval). Lloyd [32, 33] and Max [34] independently derived methods to design a quantizer subject to these conditions, which is therefore called a *Lloyd-Max quantizer*. When FRQ is used the number of cells [5] K, is related to the rate by $K = 2^r$. Evaluating expression (1.20) for a Gaussian source gives, at

[3] The first paper to compare the performance of a specific quantizer to the Shannon lower bound was that of Koshelev in 1963 [29].

[4] These assumptions are known as Bennett's assumptions [27].

[5] Once optimized, the cells are called *Voronoi regions*.

high rates,

$$\begin{aligned} d_{Gaussian,Lloyd-Max} &\approx \frac{\sqrt{3}\pi\sigma_x^2}{2K^2} \\ &\approx \frac{\sqrt{3}\pi}{2}\sigma_x^2 2^{-2r}. \end{aligned} \tag{1.21}$$

For an entropy-constrained scalar quantizer, the partition cell should be reoptimized, and high resolution analysis shows that it is optimal for each S_i to have equal length. Thus, a simple *uniform quantizer* results in the best performance for high resolution, which is

$$d_{Gaussian,ECUQ} \approx \frac{\pi e}{6}\sigma_x^2 2^{-2r}. \tag{1.22}$$

This formula is known as the Gish and Pierce [35] high rate approximation of entropy coded uniform quantization (ECUQ). Interestingly, optimal entropy constrained uniform quantization does however not result in a uniform quantizer at low rates, but again on a p.d.f. optimized quantizer [36, 37].

In any case, the number of cells is countably infinite with variable-rate coding (VRQ). If fixed-rate coding (FRQ) is used with uniform quantization, the grid of the quantization levels covers a finite range of amplitudes and the distortion is comprised of two factors. The first stems from the distortion occuring by approximating any value inside the grid by the corresponding reconstruction level, and is called the *granularity*. The second contribution comes from the distortion occuring by approximating any value outside the grid by the corresponding reconstruction level at the boundary of the grid, and it is called *overload*. For bounded uniform quantization (BUQ), the optimal stepsize depends consequently on the p.d.f. of the source and on the rate (see [14], p. 127, Table 4.1, for optimisation results w.r.t. several p.d.f.s).

Summarizing these results, the performance of a quantizer may be described by a distortion-rate function of the form

$$d \approx c2^{-2r}\sigma_x^2, \tag{1.23}$$

where σ_x^2 is the variance of the scalar source, and c is a coefficient that depends on the rate, on the p.d.f. of the source, and on the type of quantization (fixed rate, variable rate or entropy constrained). It is important to emphasize that the coefficient c of operational distortion-rate functions tends to a constant only at moderate to high rates. The Shannon Lower bound is given by

$$d_{Shannon} = 2^{-2r}\sigma_x^2, \tag{1.24}$$

The rate-distortion performance discussed above are plotted in Figure 1.3.

The distortion obtained with p.d.f. optimized FRQ (1.21) is worse by a factor of ≈ 2.7 than that of the Shannon lower bound, while the distortion of ECUQ (1.22) is $\frac{\pi e}{6} \approx 1.4$ greater than the best achievable distortion. Equivalently, for the same distortion level, FRQ requires an excess bitrate of ≈ 0.72 bit/sample as compared with the lower bound, while this excess bitrate is only ≈ 0.25 bit/sample for ECUQ. The excess bitrate of ECUQ is often quoted as the “quarter bit result”, and was first reported in [29], and rediscovered by numerical evaluation [30] in the case of i.i.d. Gaussian sources. The (slightly) subsequent paper of Gish and Pierce [35] brought then several important results. It demonstrated analytically the “quarter bit” result of ECUQ of Gaussian sources, and showed that this performance could be attainable for *any source distribution*. They also generalized the results from squared-error distortion to nondecreasing functions of magnitude error. Important analytical results relating differential (1.4) to discrete entropy (1.3) can also be

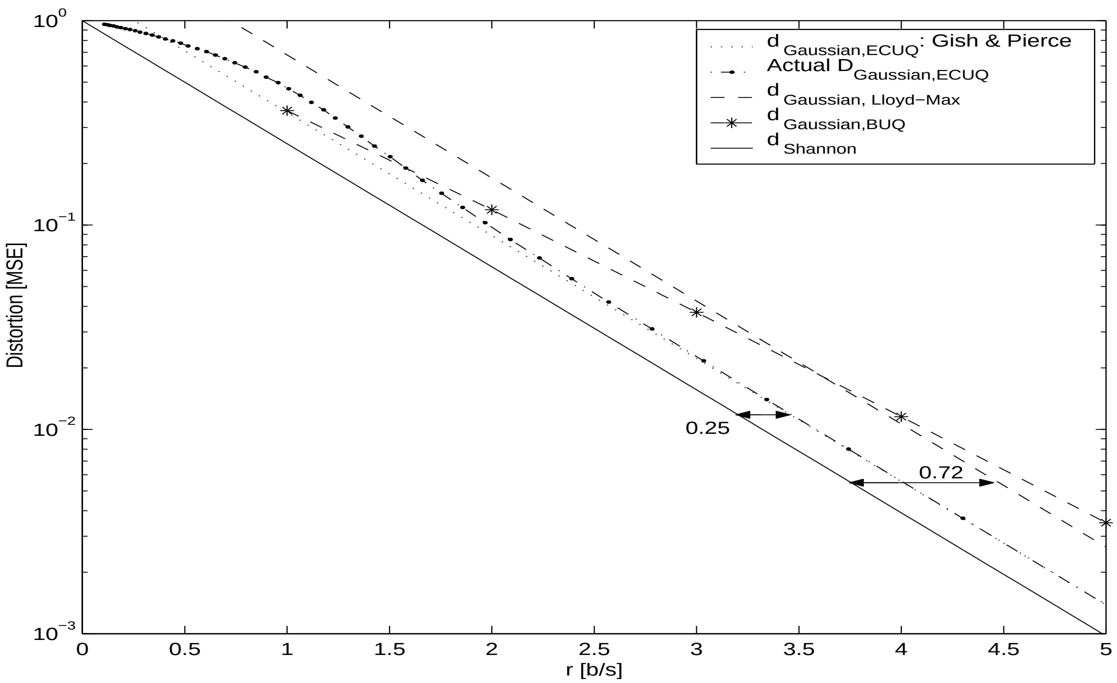


Figure 1.3: Distortion-rate functions of several scalar quantizers.

found in the same paper. A uniform quantizer scalar with infinitely many levels and small cells width Δ has output entropy given approximately by [6]

$$H(x^q) \approx h(x) - \log_2 \Delta. \tag{1.25}$$

In the high resolution case, the entropy of N successive outputs of a uniformly scalar quantized stationary source is

$$H(\underline{x}^q) = H(x_1^q, x_2^q, \cdots, x_N^q) \approx h(x_1, x_2, \cdots, x_N) - \log_2 \Delta. \tag{1.26}$$

Finally, they showed that the 0.25 bit/sample result is also true for sources with memory, and noted that when coding vectors, the performance could be improved in two dimensions by using hexagonal cells (instead of the cubic cells induced by uniform quantization).

1.5.3 Transform Coding

The first occurence of transform coding in digital systems is attributed to Huang and Schultheiss [1]. A previously introduced coding procedure aimed of transmitting linear combinations of time- and amplitude-continuous signals instead of the original signals was introduced by Kramer and Matthews [39]. The “quantization” operation corresponded in this case to transmitting a linear combination of $n < N$ signals instead of the same number of original signals, which was found to be much more efficient for adequately chosen transforms. The modern framework of transform coding introduced by Huang and Schultheiss and including digitization is depicted in Figure 1.4.

[6] This result is due to Rényi [38], and was generalized in [35] to nonuniform quantizers, and to vectors.

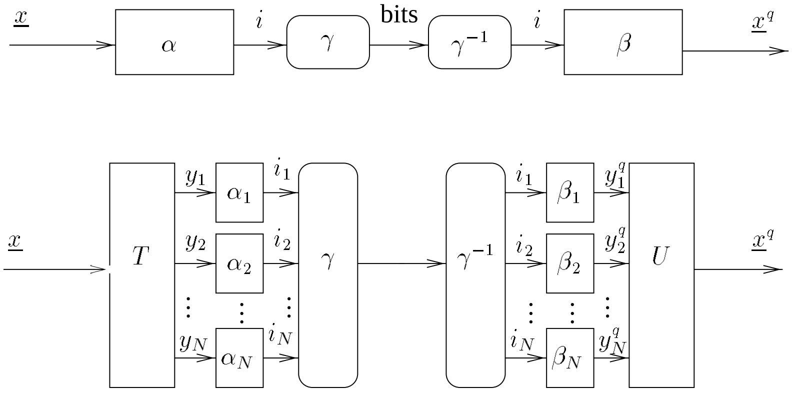


Figure 1.4: Modular structure of a transform code [1].

The essence of transform coding is the modular structure provided by breaking the mapping α into two steps, an invertible transform T producing *transform coefficients* y_i, and the independent scalar quantization of those coefficients. The corresponding serie of indexes are then compressed by an entropy code γ, which is usually itself composed of N independent entropy coders. An approximation $\underline{x}^q$ of $\underline{x}$ is then obtained by reversing the operations at the decoder [7]. The quantizers indices are first recovered, from which the decoder produces reconstructed transform signals y_i^q. The final step usually uses $U = T^{-1}$ to obtain $\underline{x}^q$. The great advantage of transform codes comes from their complexity reduction: the time for computing the transform is at most proportional to N^2, whereas computing the optimal code is exponential in N. Transform coding allows therefore large values of N to be practical, at the cost of being suboptimal. It is therefore a *constrained* code, w.r.t. its particular structure.

Bit allocation

Coding (quantizing and entropy coding) each transform coefficient separately splits the total number of bits among the transform coefficients. One should then cleverly choose the quantization fineness, and hence the number of bits required to represent the resulting quantized sources. The formulation of the bit allocation problem is simple: one is given a set of transform signals with variances $\mathrm{E}\,\widetilde{y}_i^2 = d_{i,T}$, and a set of scalar quantizers with distortion-rate performance

$$d_{i,T} = f_i(r_i), \tag{1.27}$$

where r_i are the nonnegative (and possibly noninteger) bitrates of the components y_i, and f_i describes the performance of the quantizer. The problem is to minimize the average distortion $\mathrm{E}\,||\underline{\widetilde{y}}||_T^2 = \frac{1}{N}\sum_{i=1}^{N} d_{i,T}$ subject to the constraint of a given maximum average rate $r = \frac{1}{N}\sum_{i=1}^{N} r_i$. If the average distortion can be reduced by taking bits away from one component and giving them to another, the bit allocation is not optimal. Applying this reasonning with infinitesimal changes in the component rates, a necessary condition for an optimal bit allocation is that the slopes of each f_i at r_i is equal to a common, constant value. The problem becomes easy if the operational distortion-rate function is given by (1.23) and high rate is assumed

[7] The channel plays no role in the optimization and is assumed to cause no transmission error.

(that is, $\mathrm{E}\,||\tilde{y}_i||^2 \approx c_i 2^{-2r_i}\sigma^2_{y_i}$, where c_i is some performance factor, independent of the rate). If one neglects further the fact that the r_i should be nonnegative[8], the optimal bitrates are given by

$$r_i = r + \frac{1}{2}\log_2 \frac{c_i}{(\prod_{i=1}^{N} c_i)^{\frac{1}{N}}} + \frac{1}{2}\log_2 \frac{\sigma^2_{y_i}}{\left(\prod_{i=1}^{N}\sigma^2_{y_i}\right)^{\frac{1}{N}}}, \tag{1.28}$$

and the corresponding average distortion by

$$\frac{1}{N}\,\mathrm{E}\,||\underline{\tilde{y}}||^2_T = 2^{-2r}\left(\prod_{i=1}^{N} c_i\sigma^2_{y_i}\right)^{\frac{1}{N}}. \tag{1.29}$$

If it turns out that some rates are negative, they are set to zero, and the remaining components have correspondingly higher allocations.

With average rates of r bits per component and Gaussian signals, the distortion-rate performance of the quantizers may be approximated by (1.21) or (1.22), and the average distortion with optimal bit allocation becomes

$$\frac{1}{N}\,\mathrm{E}\,||\underline{\tilde{y}}||^2_T = c2^{-2r}\left(\prod_{i=1}^{N}\sigma^2_{y_i}\right)^{\frac{1}{N}}, \tag{1.30}$$

where $c = \frac{\pi e}{6}$ for ECUQ or $c = \frac{\sqrt{3}\pi}{2}$ for optimal FRQ.

An important property of commonly used (that is, orthogonal) transformations is that, if a noise (for example quantization noise) is added to the signal in the transformed domain, then its power will be the same in the transformed and in the signal domains. This property is sometimes referred to as *unity noise gain* property. The *coding gain* G_T for a transformation T which verifies unity noise gain property is then defined as the factor by which the distortion is reduced because of the transform. Assuming high rate and optimal bit allocation

$$G_T = \frac{\mathrm{E}\,||\underline{\tilde{x}}||^2_I}{\mathrm{E}\,||\underline{\tilde{x}}||^2_T} = \frac{\mathrm{E}\,||\underline{\tilde{x}}||^2_I}{\mathrm{E}\,||\underline{\tilde{y}}||^2_T} = \frac{\left(\det\ \mathrm{diag}\,\{R_{\underline{xx}}\}\right)^{\frac{1}{N}}}{\left(\det\ \mathrm{diag}\,\{R_{\underline{yy}}\}\right)^{\frac{1}{N}}}, \tag{1.31}$$

where I is the identity matrix, and the notation $\mathrm{E}\,||\underline{\tilde{x}}||^2_T$ denotes the variance of the quantization error on the vector $\underline{x}$, obtained for a transformation T.

Optimization of the transform

Now, the problem remains of optimizing the transform so that the distortion, resulting from the bit allocation algorithm, is minimized. Firstly, among the possible choices of transforms, orthogonal transforms are traditionally prefered because they avoid a possible noise amplification when coming back into the signal

[8] This is in fact implicitly assumed in the assumption of constant c_i.

domain. Denoting by $\frac{1}{N}\,\mathrm{E}\,||\underline{\widetilde{x}}||_T^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_{\widetilde{x}_i}^2 = \dfrac{1}{N}\,\mathrm{tr}\,\{R_{\underline{\widetilde{x}\widetilde{x}}}\}$, we have for orthogonal transforms

$$\begin{aligned} \tfrac{1}{N}\,\mathrm{E}\,||\underline{\widetilde{y}}||_\perp^2 &= \tfrac{1}{N}\,\mathrm{tr}\,\{\,\mathrm{E}\,\underline{\widetilde{y}}_k\underline{\widetilde{y}}_k^T\} = \tfrac{1}{N}\,\mathrm{tr}\,\{\,\mathrm{E}\,(\underline{y}_k - \underline{y}_k^q)(\underline{y}_k - \underline{y}_k^q)^T\,\} \\ &= \tfrac{1}{N}\,\mathrm{tr}\,\{\,\mathrm{E}\,(T(\underline{x}_k - \underline{x}_k^q)(\underline{x}_k - \underline{x}_k^q)^T T^T)\} \\ &= \tfrac{1}{N}\,\mathrm{tr}\,\{T R_{\underline{\widetilde{x}\widetilde{x}}} T^T\} \\ &= \tfrac{1}{N}\,\mathrm{E}\,||\underline{\widetilde{x}}||_\perp^2 . \end{aligned} \tag{1.32}$$

Secondly, considering (1.30) or (1.31), the choice of the transform is guided by minimizing the geometric mean of the variances. Consider the covariance matrix $R_{\underline{yy}}$ of the transform signals. Since $R_{\underline{yy}}$ is positive semidefinite, it verifies Hadamard's inequality [40],

$$\prod_{i=1}^{N}\sigma_{y_i}^2 \geq \det R_{\underline{yy}}. \tag{1.33}$$

Since $\det R_{\underline{yy}} = \det(T R_{\underline{xx}} T^T)$, this determinant becomes $\det R_{\underline{xx}}$ for any orthogonal, and more generally any unimodular transform. Thus, the product in (1.33) is at least $\det R_{\underline{xx}}$, and the coding gain (1.31) of unimodular transforms is at most [41]

$$G^0 = \left(\frac{\det diag\{R_{\underline{xx}}\}}{\det R_{\underline{xx}}}\right)^{\frac{1}{N}}. \tag{1.34}$$

1.5.4 Karhunen-Loève Transform

A Karhunen-Loève Transform (KLT) is a particular type of orthogonal transform that depends on the covariance of the source. An orthogonal matrix V represents a KLT of $\underline{x}$ if $V R_{\underline{xx}} V^T$ is a diagonal matrix. This diagonal matrix is the covariance $R_{\underline{yy}}$ of $\underline{y}_k = V\underline{x}_k$. Thus, a KLT yields uncorrelated transform coefficients. KLT is the most commonly used name for these transforms in signal processing, communication and information theory, recognizing the works [42] and [43]. Among other names are Hotelling transform [44] and principal component transform.

A KLT exists for any source because covariance matrices are symmetric, and symmetric matrices are orthogonally diagonalizable. The diagonal elements of $V R_{\underline{xx}} V^T$ are the eigenvalues of $R_{\underline{xx}}$. Note that for a given source with covariance matrix $R_{\underline{xx}}$, KLTs are not unique: any row of V can be multiplied by ± 1 without changing $R_{\underline{yy}}$, and permuting the rows leaves $R_{\underline{yy}}$ diagonal.

Let us consider a jointly Gaussian source, which is tranform coded as in Fig 1.4 with $U = T^{-1}$. Since the transform coefficients have the same normalized densities, the quantizer's distortion-rate functions may be described by a single function f as $\mathrm{E}\,\widetilde{y}_i^2 = d_i = \sigma_{y_i}^2 f(r_i)$, $i = 1, \cdots, N$. Then for any bit allocation $(r_1, r_2, \cdots, r_N)$, there is a KLT that minimizes the distortion [45]. In particular, at high rates, the maximal coding gain (1.34) is achieved by the KLT. Consequently, the KLT is often used as a benchmark in transform coding.

However, if neither the Gaussian, nor the $U = T^{-1}$ assumptions are valid, there are cases where the KLT is not an optimal transform. The optimality of the KLT for transform coding of Gaussian sources is believed to be a consequence of the fact that the KLT of a Gaussian vector yields independent transform coefficients. The application of the KLT in transform coding of non-Gaussian sources is then justified using the intuition that the KLT's coefficient decorrelation is, for general sources, the best possible approximation to the desired coefficient independence. The successes and failures of this intuition are reviewed in [46]. Several cases where the KLT is not optimal are described in [47]. Other recent works include [48, 49, 50]. A large part of the present work deals with the performance of a new transform, namely the causal transform. The proposed investigations are thus premiliminary, and the Gaussian assumption allows one to set a framework in which analytical results can be derived. Therefore, we will not enter the details of the analyses related to general sources; the KLT will be considered as the optimal transform in this thesis.

1.6 Thesis Themes and Overview

The thesis is comprised of two parts. The first one deals with lossy coding, and the second one with lossless coding. A brief overview of the general framework of this thesis, and of each part is given in this section. More detailed introductions to the specific frameworks considered for lossy and lossless coding can be found at the beginning of each part; an abstract is provided at the beginning of each chapter.

The topic of causality in source coding is the essential link between the several chapters of this thesis. Several causal decorrelating schemes will be, somewhat paradigmatically, investigated. In all the cases where the considered causal coding scheme has the form of a (scalar valued) matricial transform, comparison will be made with the Karhunen-Loève transform. Inspiring from [19] and [20], this thesis could also have been titled “Variations on a causal coding theme”; however, an effort was made so that the chapters can be read independently. We tried to briefly but clearly recall the previously established background and results, whenever it seemed necessary.

The *information* considered in this thesis will be represented by vectorial signals (whose samples are vectors), which compose a wide class of signals, among which scalar and multichannel signals. Multichannel signals may be obtained as soon as scalar signals are, in the context of various applications, gathered together. If these signals present some dependencies, such as audio signals for example, one may process them jointly in order to achieve a more efficient compression.

The initial idea of developping coding techniques for audio signals[9] motivated this choice of a vectorial representation. Though some applications will be presented for this kind of signals, Gaussian source models is often assumed. Gaussian sources have indeed a particular status in information theory. Shannon [25] showed that a Gaussian i.i.d. source has the worst rate-distortion function of any i.i.d. source with the same variance, thereby showing that the Gaussian source is an extremum in a source coding sense. This fact provided an approach to *robust quantization*: the resulting code might not be optimal for the actual source, but would perform no worse than it would on the Gaussian source for which it was designed (see *e.g.* [51]). Besides, advantage can be taken of the central limit theorem and of the known structure of an optimal quantizer for a Gaussian random variable. A general source is in this case coded by first filtering it to produce approximately Gaussian density, scalar quantizing the result, and then inverse filtering to recover the quantized original [52]. We will not argue however that Gaussian assumption was intended to provide either worst-case performance, nor methods to code, with the same performance as that obtained in the Gaussian case, sources with arbitrary densities. The Gaussian source model was retained in this work

[9] The first results of this work were obtained in the framework of the french RNRT project *COBASCA* :*CO*dage en *B*ande élargie avec partage *A*daptatif du débit entre *S*ource et *CA*nal pour réseaux cellulaires de deuxième et troisième générations (UMTS).

because it allows one to derive closed form expressions, to compare, and possibly to prove the optimality of the considered coding schemes. In this sense, it provides a valuable framework for preliminary theoretical investigations, such as those intended in this work.

1.6.1 Part One: Causal Lossy Coding

The first part deals mainly with transform coding. Transform coding theory may appear as rather old and routine; a primary aim of this thesis is to show that valuable innovations are still possible. These innovations regard not only the framework of the standard description of transform coding (by introducing a new causal transform and showing its efficiency for a wide range of rates), but also in the framework of a related and almost unexplored research areas, namely the problem of *backward adaptation* in transform coding.

The causal transform is introduced in chapter 2. It is called LDU transform, for “Lower-Diagonal-Upper” factorization of the correlation matrix of the input vectorial source; the matrix is lower triangular and unit diagonal, and its design is based on optimal prediction. Both theoretical analyses and empirical evidence of its coding performance w.r.t. the KLT are demonstrated. The presented theoretical investigations apply to the classical high rate transform coding framework, but particular practical systems working at moderate to low rates are also investigated.

Because it is based on optimal prediction, the LDU is, as the KLT, signal dependent. Adjusting the transform to the generally varying changing covariance matrix of the source may result for practical systems in a non acceptable bitrate overhead. A possible way to avoid this drawback is to adapt the transform based on the decoded data, so that the encoder and the decoder adapt in unison without the explicit transmission of any coding parameters. Among the questions of interest is that of knowing whether the backward adaptive system will be suboptimal, in the rate-distortion sense, w.r.t. to a system designed with a perfect knowledge of the source. This issue is addressed in the third and fourth chapters.

In a first step, the approach of chapter 3 makes the same assumptions as those of the classical transform coding framework (optimal bit allocation procedure, high rate, Gaussian sources), and proposes an analysis of the coding gain for such an idealized backward adaptive system. The approach consists in modelling the behaviour of these systems by considering the effects of quantization and estimation noise as perturbation terms on the ideal classical transform coding framework. The perturbation effects impacts both the bit assignment mechanism, and the transforms. The analyses are made in both the causal and unitary cases.

The previous approach assumes however an optimal bit assignment mechanism which may not be the case for practical systems. Therefore, an analysis of practical algorithms is proposed in chapter 4, for which the optimal bit allocation assumption is released, and replaced by a simple (equal stepsize) quantization rule. Both constant and adaptive stepsizes are considered, though emphasis is put on algorithms using adaptive stepsizes, in order to cope with possible variations of the energy of the sources.

These topics are followed by a generalization of the presented causal coding scheme to (matricial) filtering in chapter 5. For vectorial sources with memory, instantaneous decorrelation such as that performed by transforms such as KLT or LDU applied to vector samples is not optimal. Temporal redundancies may remain, which will not be accounted for by scalar entropy coders.

By considering blocks of vectors with infinite length, we show that the optimal decorrelating approach is still lower triangular. The scalar coefficients of the LDU matrix are in this case replaced by prediction filters. We show that this renders the coding procedure of Gaussian vectorial sources with memory optimal (assuming high rate and filters of infinite length). This generalization of the LDU is called "generalized Multiple-Input/Multiple Output (MIMO) prediction". This approach includes, as special cases, previously introduced MIMO decorrelation approaches, and turns out to be rich of both theoretical and practical consequences[10]. A high rate analysis provides an optimal ordering in the decorrelation of the signals, and gives insight about which particular decorrelation approach should be prefered to make the coding scheme the most efficient.

A brief history of the preliminary analyses, and the framework which led to these results are then presented in an appendix chapter.

1.6.2 Part Two: Causal Lossless Coding

The second part of this thesis presents and analyzes lossless coding techniques based on the causal decorrelating approaches described in the chapters 2, 3 and 4. The analysis of the performance of the LDU transform in a lossless coding framework was first motivated by our interest in pursuing the comparison of its coding performance with the KLT, with the intuition that the LDU might, due to its simple triangular structure, outperform in this framework its othogonal brethren. Also, multichannel lossless audio coding has recently become a challenging field, and the results of chapter 5 may inherently be applied to lossless coding of multichannel sources. Besides audio, the results presented in the second part may also be applied to the field of image coding.

Basically, the structure of this second part resembles that of the first one. In the first two chapters, the LDU causal transform is compared to orthogonal transforms, and in particular to the KLT, in a lossless transform coding framework. The last chapter investigates the extention of the LDU transform, or generalized MIMO prediction, to optimal lossless coding of vectorial signals.

An ubiquitous topic in the second part is that of integer-to-integer transforms, which received much attention recently. The term comes the fact that both the inputs and the outputs of these transforms are integer valued (or lie on a scaled integer lattice). These transforms are thus of interest in a lossless coding framework, where they can be applied to discrete-amplitude source, such as those resulting from some quantization process. The corresponding systems will be denoted by "single-stage", or "one-shot" lossless coders. The goal of the integer-to-integer transforms is to provide systems which present (almost) the same compression performance as those obtained by vectors entropy coders, though using scalar entropy codes. A particular emphasis will be put on the "almost" of the last sentence. Theoretical analyses will first evaluate the suboptimality of realizable integer-to-integer structures followed by scalar entropy codes, w.r.t. optimal vector entropy coding methods. The analyses of these structures (and those of the corresponding bounds) will regard the integer-to-integer implementations of the KLT and of the LDU in chapter 6. In chapter 8, we will investigate those of the MIMO decorrelation approaches described in chapter 5.

[10] Beyond the field of source coding, generalized MIMO prediction has also found a natural and usefull application to multiuser detection [53].

Besides these “one-shot” or “single-stage” approaches, another recurrent theme of this second part is that of multiresolution lossless coding. These systems aim of providing a low resolution (lossy coded) version of the signals in a first step; the error signals are transmitted in a second step. For these systems, it appears interesting to know whether the multiresolution approach is suboptimal w.r.t. the corresponding single-stage system. We will therefore analyze the bitrates dedicated to code the low resolution, and the error signals, and compare the resulting overall bitrate to that obtained with single-stage coders. These comparisons will be done for the LDU transform and the orthogonal transforms in chapter 7, and for the classical and triangular MIMO predictors in chapter 8.

Part I

Causal Lossy Coding

Overview of the First Part

Transform codes are popular because they provide an attractive compromise between computational complexity and performance. As summarized in the introdution of this thesis, this technique has been widely analyzed, and source coding systems which use transform codes are ubiquitous. Many transforms exist, which allow different trade-offs between the theoretical coding efficiency and more practical criteria. Theoretical coding efficiency include decorrelation efficiency, or compaction gain. More practical criteria include design or implementation complexity, or subjective performance related to the particular behaviour of the transforms w.r.t the nature of the signals they are applied to. A pervasive use is made of orthogonal transforms, since they guarantee that the quantization noise will not be amplified when coming back from the transform to the signal domain. Among them, the Karhunen-Loève transform has become a benchmark, since it has been proven to be optimal for Gaussian sources [1, 54]. A recurent theme in this thesis is to show that, w.r.t. different criteria, KLT's performance may be approached, achieved, or even surpassed by another transform, namely the *causal* transform.

- In chapter 2, we will introduce the proposed transform coding technique, which is based on optimal prediction. The corresponding transform performs a Lower-Diagonal-Upper factorization of the covariance matrix of the vectorial source to be coded. It is not unitary but causal: the transform matrix is unit diagonal and lower triangular. A theoretical analysis shows first that at high rates, it may achieve the same peformance as the KLT. As a consequence of its non-orthogonality, we show that efficient causal coding structures should be implemented in closed loop around the quantizers, as in DPCM systems. As a consequence of the closed loop implementation, a noise feedback should increase the resulting distortion at lower rates. The point is then to know quantitatively how the noise feedback impacts the coding performance. We propose therefore theoretical analyses of the noise feedback. In a first step, general quantizers, high rate and optimal bit assignment are assumed. In a second step, the performance of practical systems using nearly optimal by allocation, (uniform quantization with equal quantization stepsize, and entropy coding) are evaluated. Both theoretical analyses and numerical results will show that the causal transform competes with the KLT at average bitrate budgets as low as 2.5 b/s. These results were presented in [55, 56].

  As the KLT, the LDU is data dependent, and should thus be updated in case of changes in the source statistics. In order to avoid transmitting coding parameters as side information, one may attempt to adapt the transforms using decoded data only. This poses the problem of *backward adaptation*, or *adaptation without side-information*, or *on-line adaptation* in transform coding. The feasibility and performance analysis of this kind of coding schemes will be the topics of the following two chapters.

- Chapter 3 presents a first attempt to model theoretically the performance of causal and unitary backward adaptive coding schemes. The proposed approach will consist in analyzing the perturbation effects w.r.t. to the ideal case of the classical transform coding framework, where the second order statistics of the source are known. In order to make tractable analyses, several simplifying assumptions are made, which are borrowed from the classical high rate transform coding framework. Namely, we assume Gaussianity, same quantizers'rate-distortion law, and bit assignment rule of the form (1.28), in which however the actual variances of the transform signals are not known. The proposed model accounts then for perturbations occuring uppon both the bit assignment mechanism and the transforms' design. Three cases will be investigated: the coding schemes are perturbed by quantization noise only in a first case, and by estimation noise only in a second case; finally, both effects will be accounted for. Theroretical evaluations will be shown to describe correctly this kind of systems. These results we presented in [57, 58, 59].

  Our goal to provide a successful analysis of backward adaptive transform coding schemes seems however somewhat incomplete at this point: practical systems may not verify the above assumptions. This leads to the topics of chapter 4.

- In chapter 4, three practical backward adaptive transform coding schemes will be investigated for both the causal and the unitary transforms. In these algorithms, the quantization stepsizes are the same for all the transform components; the transforms are computed using estimates of the covariance matrices based on quantized data. In a first step, constant (w.r.t. time) stepsize algorithms are implemented. This case is of interest if the input source is stationary; otherwise, it may result in unacceptable changes in the rate-distortion performance. The point is to know whether the transforms will converge or not to optimal transforms (designed with the knowledge of the statistics of the original source). Empirical evidence will show that this is the case, even at low rates. In a second step, we propose a theoretical analysis of two algorithms using adaptive stepsizes. The adaptation procedure is similar to that used in classical adaptive scalar quantization. We model then the expected distortion obtained for a given number of decoded vectors. Our results suggest convergence of both the stepsize and the transforms, for both algorithms. In the case where the source is stationary, the algorithm using a Sheppard's correction on the second order moment estimates allows one to reach a target point of the rate-distortion function of the system. This point is reached by the structure after a convergence process, though the decoder has *a priori* neither the knowledge of the stepsize to be used, nor that of the statistics of the source. These results are presented in [60].

  The causal transform studied in these first chapters proves efficient decorrelation ability. As the KLT however, it accounts only for correlations *within* each data block. For vectorial sources whose vectors are not independent, better coding efficiency can be expected from tranforms which account for temporal redundancies as well. This is the topic of the last chapter of this first part, which generalizes in this sense the causal approach investigated so far.

- We show in chapter 5 how the causal transform LDU can be extended to (matricial) filtering. In this case, the optimal causal decorrelating scheme will be shown to correspond to a triangular prediction matrix whose entries are optimal prediction filters. The diagonal filters are scalar intrasignal prediction filters, and the off-diagonal predictors are Wiener filters performing the intersignal decorrelation. By considering vectors of infinite size, one can get frequential expressions for the coding gains. We show that this decorrelating scheme leads to the notion of *generalized* MIMO (Multiple Input Multiple Output) prediction, in which a certain degree of non causality may be allowed for the off-diagonal prediction filters. Previously introduced MIMO decorrelation approaches are shown to appear as special cases of the described decorrelation technique.

In the case of non causal intersignal filters, the optimal MIMO predictor is still triangular, and hence “causal, in a wider sense. The notion of causality may be generalized: the causality between channels becomes processing the channels in a certain order. Some signals may be coded using the coded/decoded versions of the “previous” signals. An interesting result is that if the quantization noise feedback is taken into account, the triangular predictor is the more efficient. Moreover, the coding gain is maximized if the signals are decorrelated by order of decreasing variance. These results were presented in [61].

Chapter 2

Optimal Causal versus Unitary Transform Coding

In a transform coding framework, we introduce the optimal (linear) decorrelating transform subject to the constraint of causality. This transform is shown to correspond to a Lower-Diagonal-Upper (LDU) factorization of the signal covariance matrix $R_{\underline{xx}}$. The LDU transform is compared to the unitary approach (Karhunen-Loève Transform, KLT), which is optimal for Gaussian sources. The performance of the LDU transform is first shown to be equivalent to that of the KLT at high rates. Moreover, it presents several advantages w.r.t. its unitary counterpart, such as lower implementation and design complexities, and perfect reconstruction property. As in classical (A)DPCM, closed loop implementation of the causal coding structure is shown to be preferable. This leads to a noise feedback effect, similar to that occuring in DPCM systems. We present high resolution analysis of these effects on the distortion-rate function. The proposed analyses consider firstly general transform coding systems for which the bit allocation is optimal, and secondly practical systems whose bit allocation is nearly optimal. For the latter system, deviations from high rate assumptions arise approximately beyond 3 b/s. The effects of the noise feedback in the causal case become non negligible below approximately 2 b/s. The theoretic evaluations are validated by numerical results.

2.1 Introduction

The unitary Karhunen-Loève Transform is known to be the optimal transform for Gaussian[1] sources [54] This chapter analyzes the performance of the optimal *causal* transformation for Gaussian signals for variable length coding at high and low bitrates .

Rather surprisingly regarding its excellent coding performance, this transform was derived only recently (independently in [62, 63] [55], and [64]). In [62], the transform is named *Prediction based Lower triangular Transform* (PLT). The work [64] calls the transform *Sequential Vector Decorrelation Technique*. In [55], the causal decorrelation approach was named VDPCM because it generalizes scalar DPCM. The term will not be retained here in order to avoid confusion with the technique presented in [65], which uses vector quantization (see also section 5.1).

The causal transform is described in the second section, and the analogy with scalar DPCM is underlined, as its coding performance depends on wether it is implemented in open or in closed loop around the quantizers. In the third part, the coding performance with negligible feedback is analyzed, and further comparisons with the KLT are presented in section 2.4. Since the noise feedback arises in actual implementable causal coding structures, a realistic analysis of the coding performance of this transform should evaluate what, in terms of rate, does coarse quantization correspond to, and how the corresponding coding performance is actually deteriorated. No such analyses were proposed in [62, 63] nor in [64]; this is the aim of the sections 2.5 and 2.6. Section 2.5 proposes an analysis based on high rate and optimal bit allocation assumptions, and a practical system is investigated in section2.6. The last part summarizes the main results and draws some conclusions.

2.2 Causal Transform Coding

2.2.1 Open Loop and Closed Loop Causal Transform Coding

Let us consider the coding scheme of figure 2.1. A matrix transformation L is applied to the vector $\underline{x}_k = [x_{1,k} \cdots x_{N,k}]^T$: $\underline{y}_k = L\underline{x}_k = \underline{x}_k - \overline{L}\underline{x}_k$, where $\overline{L}\underline{x}_k$ is the reference vector. The difference vector $\underline{y}_k = [y_{1,k} \cdots y_{N,k}]^T$ is then quantized using a set $\mathbf{Q}$ of (variable- or fixed- rate) scalar quantizers Q_i. The output $\underline{x}_k^q$ is then $\underline{y}_k^q + \overline{L}\underline{x}_k$. This scheme may be considered as a generalization of the scalar (open loop) DPCM coding scheme.

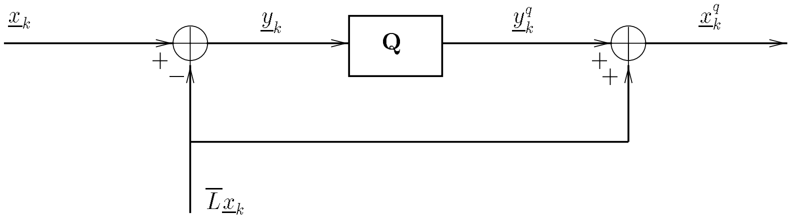


Figure 2.1: Open loop causal transform coding ($\mathbf{Q}$ denotes a set of scalar quantizers).

[1] For non Gaussian sources, different transforms may yield better compression results, see e.g. [47, 49, 50].

As in scalar DPCM, the reconstruction error vector $\widetilde{\underline{x}}_k$ equals the quantization error vector $\widetilde{\underline{y}}_k$ since

$$\widetilde{\underline{x}}_k = \underline{x}_k - \underline{x}_k^q = \underline{x}_k - (\underline{y}_k^q + \overline{L}\underline{x}_k) = \underline{x}_k - \overline{L}\underline{x}_k - \underline{y}_k^q = \underline{y}_k - \underline{y}_k^q = \widetilde{\underline{y}}_k. \tag{2.1}$$

Note that L behaves as the Identity matrix w.r.t. the noise vector $\widetilde{\underline{y}}_k$ introduced in the transform domain. As in the unitary case, the power of the quantization noise is thus the same in signal and transform domains[2]. Since the constraint imposed here on the linear transformation is causality, the matrix $\overline{L} = I - L$ is strictly lower triangular. The nonzero elements of $\overline{L}$ represent the degrees of freedom of the transformation. The causality refers to the ordering of the signals x_i which compose $\underline{x}$. This notion could be generalized by working with the permuted components of $\underline{x}$ and $\underline{y}$, which gives $\mathcal{P}\underline{y} = L\ \mathcal{P}\,\underline{x}$ or $\underline{y} = (\mathcal{P}^T\ L\mathcal{P})\underline{x}$, where $\mathcal{P}$ is a permutation matrix. This will be developped in chapter 5.

As in an open loop DPCM coding scheme however, the coding system represented by figure 2.1 suggests that not only information concerning the prediction residual should be transmitted to the decoder, but also an accurate version of every reference vector, which from a bitrate point of view is not realistic. If on the other hand a reference vector $\overline{L}\underline{x}_k$ is used at the encoder, and a different $\overline{L}\underline{x}_k^q$ at the decoder, the system would suffer from quantization noise amplification, which may unacceptably decrease the coding performance, or even make the prediction structure useless. The closed loop coding scheme of figure 2.2 will therefore be prefered. In this case, the reference signal $\overline{L}x_k^q$ is based on the past quantized samples (available at both the encoder and the decoder).

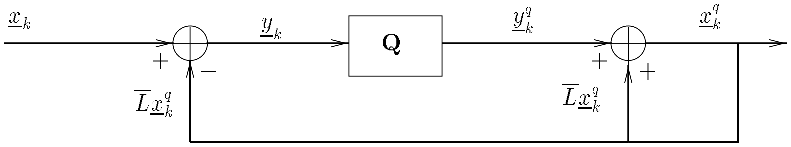


Figure 2.2: Closed loop causal transform coding ($\mathbf{Q}$ denotes a set of scalar quantizers).

In this case, reconstruction and quantization errors are still equal, since

$$\widetilde{\underline{x}}_k = \underline{x}_k - \underline{x}_k^q = \underline{x}_k - (\underline{y}_k^q + \overline{L}\underline{x}_k^q) = \underline{x}_k - \overline{L}\underline{x}_k^q - \underline{y}_k^q = \underline{y}_k - \underline{y}_k^q = \widetilde{\underline{y}}_k. \tag{2.2}$$

Two particular implementations of the closed loop causal transform will be reviewed in section 2.5.

In a first step (section 2.3), we neglect the quantization error on the reference signal. The coding performance of a closed loop causal transform with non negligible feedback will then be described in section 2.5. In any case, we will suppose an optimal bit assignment and make high resolution assumptions. A practical analysis at lower rates, in the case of a nearly optimal bit assignment is presented in the last section. Moreover, one assumes jointly Gaussian r.v.s x_i, whith known covariance matrix $R_{\underline{x}\underline{x}}$.

[2] In the causal case however, the transform does not only conserve the Euclidian norm of the noise, but also the shape of its p.d.f.. This property will be used in chapter 8 in the framework of multiresolution lossless tranform coding.

2.2.2 Problem statement

According to the so-called unity noise gain property (2.1), the coding gain for such a transformation L is then

$$G_L = \frac{\mathrm{E}\,||\underline{\tilde{x}}_k||_I^2}{\mathrm{E}\,||\underline{\tilde{x}}_k||_L^2} = \frac{\mathrm{E}\,||\underline{\tilde{x}}_k||_I^2}{\mathrm{E}\,||\underline{\tilde{y}}_k||_L^2}, \tag{2.3}$$

where I is the Identity matrix (which corresponds to the absence of transformation), and the notation $E||\widetilde{\underline{x}_k}||_L^2$ denotes the variance of the quantization error on the vector $\underline{x}$, obtained for the transformation L. Similarly, the SNR [3] obtained by using L may be defined as

$$\mathrm{SNR}_L = 10\log_{10}\frac{\mathrm{E}\,||\underline{x}_k||^2}{\mathrm{E}\,||\underline{\tilde{x}}_k||_L^2} = 10\log_{10}\frac{\mathrm{E}\,||\underline{x}_k||^2}{\mathrm{E}\,||\underline{\tilde{y}}_k||_L^2} = 10\log_{10}\frac{\mathrm{E}\,||\underline{x}_k||^2}{\mathrm{E}\,||\underline{y}_k||_L^2}\,\frac{\mathrm{E}\,||\underline{y}_k||_L^2}{\mathrm{E}\,||\underline{\tilde{y}}_k||_L^2} \tag{2.4}$$

We now set out to characterize the optimal transformation L and bit assignment which maximizes the coding gain. For a given bit assignment, the optimal causal transformation is

$$L = \arg\max_L G_L = \arg\max_L \mathrm{SNR}_L = \arg\min_L \mathrm{E}\,||\underline{\tilde{y}}_k||_L^2. \tag{2.5}$$

2.3 Optimal Causal Transform Coding with Negligible Feedback

In the following, the time index will be omitted in order to put emphasis on the index of the component (subscript $_i$). We assume in this section high resolution rate distortion function $\sigma_{\tilde{y}_i}^2 = \mathrm{E}\,||\tilde{y}_i||^2 = c2^{-2r_i}\sigma_{y_i}^2$ for all the quantizers. The coefficient c describes the quantizer performance; it is independent of r at high rates (e.g. $\frac{\pi e}{6}$ for ECUQ, or $\frac{\sqrt{3}\pi}{2}$ for optimal FRQ, see section 1.5.2). For a given L, the optimal bit assignment minimizes $\mathrm{E}\,||\underline{\tilde{y}}||_L^2 = \sum_{i=1}^{N}\sigma_{y_i}^2 c2^{-2r_i}$, subject to the constraint $\sum_{i=1}^{N} r_i = Nr$, where r is the average bitrate budget. The quantizers Q_i are assumed to introduce independent white noises $\tilde{y}_i$ on the components y_i, of variances $\sigma_{\tilde{y}_i}^2$ The result of this (Lagrangian) optimization yields (see section 1.5.3)

$$\sigma_{q_i}^2 = \sigma_{\tilde{y}_i}^2 = c\,2^{-2r_i}\sigma_{y_i}^2 = c\,2^{-2r}\left(\prod_{i=1}^{N}\sigma_{y_i}^2\right)^{\frac{1}{N}} = \sigma_q^2. \tag{2.6}$$

The optimal quantization error variances $\sigma_{q_i}^2$ are equal (independent of i).

Concerning the optimization of L, one should now minimize $(\prod_{i=1}^{N}\sigma_{y_i}^2)^{\frac{1}{N}}$, where the $\sigma_{y_i}^2$ depend on the rows L_i of L. The problem is hence separable, and minimizing $(\prod_{i=1}^{N}\sigma_{y_i}^2)^{\frac{1}{N}}$ with respect to L entails minimizing $\sigma_{y_i}^2$ with respect to $L_{i,1:i-1}$. The components y_i appear clearly as the prediction errors of x_i with respect to the past values of $\underline{x}$, the $\underline{x}_{1:i-1}$, and the optimal coefficients $-L_{i,1:i-1}$ as the optimal linear prediction

[3] Signal to (quantization) Noise Ratio

coefficients. The linear causal transform which minimizes (2.5) is therefore of the form

$$L = \begin{bmatrix} 1 & & & \\ \star & \ddots & \mathbf{0} & \\ \vdots & \ddots & \ddots & \\ \star & \cdots & \star & 1 \end{bmatrix},$$

where the $\star$ represent optimal prediction coefficients. In other words, L is such that

$$L R_{\underline{xx}} L^T = R_{\underline{yy}} = \overline{\text{diag}}\,\{\sigma_{y_1}^2 \cdots \sigma_{y_N}^2\}, \tag{2.7}$$

where $\overline{\text{diag}}\,\{\underline{a}\}$ represent the diagonal matrix with diagonal $\underline{a}$. Since each prediction error y_i is orthogonal to the subspaces generated by the $\underline{x}_{1:i-1}$, the transform coefficients y_i are orthogonal, and $R_{\underline{yy}}$ is diagonal. It follows that

$$R_{\underline{xx}} = L^{-1} R_{\underline{yy}} L^{-T}, \tag{2.8}$$

which represents the LDU factorization of $R_{\underline{xx}}$. Since the covariance matrix $R_{\underline{xx}}$ is positive definite, the transform L always exists. Moreover, it is unique (see Appendix 2.B).

The distortion (2.6) under high rate and optimal bit allocation becomes

$$\text{E}\,\|\underline{\tilde{y}}\|_L^2 = c\;2^{-2r}\left(\det R_{\underline{xx}}\right)^{\frac{1}{N}}, \tag{2.9}$$

and referring to (2.3), the coding gain can be written as

$$G_L^{(0)} = \left(\frac{\det\left[\,\text{diag}\,\{R_{\underline{xx}}\}\right]}{\det\left[\,\text{diag}\,\{L R_{\underline{xx}} L^T\}\right]}\right)^{\frac{1}{N}} = \left(\frac{\det\left[\,\text{diag}\,\{R_{\underline{xx}}\}\right]}{\det R_{\underline{xx}}}\right)^{\frac{1}{N}} = \left(\frac{\det\left[\,\text{diag}\,\{R_{\underline{xx}}\}\right]}{\det \Lambda}\right)^{\frac{1}{N}} = G_V^{(0)}, \tag{2.10}$$

where the superscript (0) refers to the ideal case where the quantizers have same and constant performance factor c, the rate is sufficiently high and the bit assignment optimal. The notation $\text{diag}\,\{A\}$ denotes the diagonal matrix with same diagonal as A, V denotes a KLT of $R_{\underline{xx}}$ and Λ the corresponding matrix of eigenvalues. The second and third equalities in equation (2.10) follow from the unimodularity property of L and V: both the product of the eigenvalues and that of of the prediction error variances corresponding to a covariance matrix $R_{\underline{xx}}$ equal its determinant.

Summarizing, for an optimal bit allocation, the high rate coding gains of the KLT and the LDU are the same without perturbation for three reasons : both transformations ensure that the power of the quantization error is the same in the transform and in the signal domains, they are totally decorrelating transforms, and finally they are unimodular. Moreover, this is the best coding gain achievable among all unimodular transform[4].

[4] A proof, based on Hadamard's inequality for symmetric positive semidefinite matrices may be found in [41].

2.4 Further Comparisons between Unitary and Causal Approaches

2.4.1 Complexity of the Design of an LDU transform

The optimal prediction coefficients can be solved by using the Levinson algorithm [66]. For each recursion, corresponding the a predictor of order n, this algorithm requires approximately $2n$ multiplications and a similar amount of additions. This yields a design complexity of $\sum_{i=1}^{N} 2n = 2\frac{N(N+1)}{2} \approx N^2$.

2.4.2 Complexity of the Implementation of an LDU Transform

Since N is a lower triangular and unit diagonal matrix, computing each transform N-vector $\underline{y}_k$ requires $\frac{N(N-1)}{2}$ multiplications and additions, which is less than one half the complexity required by the KLT (which requires N^2 multiplications and $N(N-1)$ additions) [67].

In the special case of AR(p) processes, the lower left corner of L will contains zeros if $N > p+1$ (one zeros will appear at the first entry of the $(p+2)$nd row, two zeros at the first two entries of the $(p+3)$th row, etc). The total complexity will therefore be reduced by $\frac{(N-p)(N-p-1)}{2}$ multiplications and additions. For an AR(1) process, $N-1$ multiplications and additions remain. The complexity of the inverse transform is indeed the same.

2.4.3 Quantization of the Coefficients

Suppose that we quantize the coefficients of the optimal causal and unitary transforms L and V, resulting in transforms L^q and V^q. On the one hand, the quantized KLT V^q will then loose its perfect reconstruction property, since $V^q V^{qT}$ will in general be different from the Identity matrix. The recovered vector is then $\hat{\underline{x}}_k = V^{qT}\underline{y}_k = V^{qT}V^q\underline{x}_k \neq \underline{x}_k$. In the causal case on the other hand, the exact vector $\underline{x}_k$ can be recovered however coarse the quantization, since $\hat{\underline{x}}_k = \underline{y}_k + L^q\underline{x}_k = \underline{x}_k - L^q\underline{x}_k + L^q\underline{x}_k = \underline{x}_k$. This means also that if the transformation coefficients are transmitted to the decoder, in a forward adaptive transform coding framework for example, the unity noise gain property (sec. 1.5.3) $\mathrm{E}\,||\tilde{\underline{x}}_k||^2 = \mathrm{E}\,||\tilde{\underline{y}}_k||^2$ will not hold anymore for the KLT.

2.5 Performance of a Closed Loop Causal Transform Coding Scheme

We first recall the results of the classical analysis regarding the noise feedback in closed loop DPCM coding schemes [14]. In all the presented analyses, ECUQ is assumed. For this type of quantizers, the additive quantization noise model is accurate for a wide range of rates (see 2.A). The operational distortion-rate functions of the quantizers are then denoted by $d_i = c2^{-2r_i}\sigma_{y_i}^2$, where c generally depends on r. For sufficiently high rates, c tends to $\pi e/6$ (≈ 3 b/s, see the numerical results in section 2.5.1) which is known as the Gish and Pierce [35] approximation. Note that the analysis of noise feedback in DPCM does not require ECUQ. Examples of DPCM systems using fixed-rate, p.d.f. optimized bounded uniform quantizers are presented in [14]. Efficient transform coding systems require however proper bit allocations; noninteger rates such as those simply provided by ECUQ are therefore more usefull, in this framework, than those obtained by fixed rate quantization. The analysis of noise feedback in scalar DPCM is then generalized to the causal transform introduced above. A high resolution analysis assuming classical optimal bit allocation is first exposed in 2.5.2. The analysis of a practical case is presented in 2.6.

2.5.1 Quantization Noise Feedback in scalar DPCM

Assume in a first step that we use a first order predictor, and the the prediction is not based on quantized data, that is, $y_k = x_k - \widehat{x}_k = x_k - \bar{l}x_{k-1}$. The reconstructed sample at the decoder is then $x_k^q = y_k^q + \widehat{x}'_k = y_k^q + \bar{l}x_{k-1}^q$.

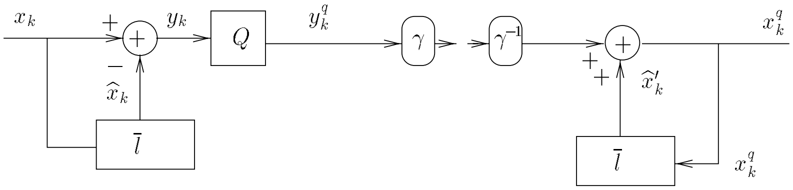


Figure 2.3: Open loop scalar DPCM coding scheme.

If we assume the process to be a first order autoregressive process with normalized correlation coefficient ρ, the optimal predictor $\bar{l}$ equals ρ, and the variance of the optimal prediction error is $\sigma_y^2 = \sigma_x^2(1-\rho^2)$. Denoting now by $\widetilde{y} = y - y^q$ and $\widetilde{x} = x - x^q$ the quantization and the reconstruction noise respectively, we obtain

$$\begin{aligned} x_k^q &= y_k^q + \rho x_{k-1}^q \\ &= x_k - \rho x_{k-1} - \widetilde{y}_k + \rho x_{k-1} + \rho\widetilde{x}_{k-1} = x_k - \widetilde{x}_k \end{aligned} \tag{2.11}$$

from which we get

$$\widetilde{x}_k = \widetilde{y}_k - \rho\widetilde{x}_{k-1}. \tag{2.12}$$

The reconstruction error differs from the quantization error. From the previous expression, their respective variances may be related by

$$\mathrm{E}\,\widetilde{x}_k^2 = \frac{\mathrm{E}\,\widetilde{y}_k^2}{1-\rho^2}. \tag{2.13}$$

Hence, what is gained in prediction is lost because the quantization noise is amplified at the decoder. The distortion-rate function of the prediction error signal is then

$$d_y(r) = \mathrm{E}\,\widetilde{y}_k^2 = c2^{-2r}\sigma_y^2, \tag{2.14}$$

and that of the overall system is

$$\mathrm{E}\,\widetilde{x}_k^2 = c2^{-2r}\frac{\sigma_y^2}{1-\rho^2} = c2^{-2r}\sigma_x^2, \tag{2.15}$$

which corresponds to a direct quantization and entropy coding of the original source x. In other words, the prediction operation is strictly useless in this system. In order to avoid quantization noise amplification, both the encoder and the decoder should use quantized data, since only quantized samples are available at the decoder side. A closed loop scalar DPCM coding scheme is presented in figure 2.4.

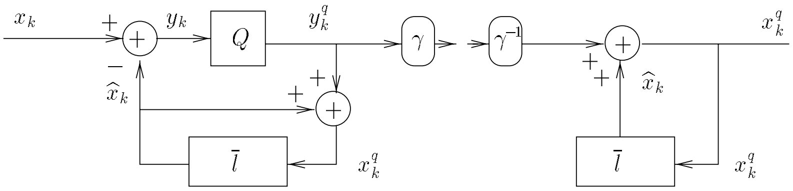


Figure 2.4: Closed loop scalar DPCM coding scheme.

The power of the quantization noise for the signal and for the prediction residual are then equal since

$$\mathrm{E}\,\widetilde{x}_k^2 = \mathrm{E}\,(x_k - x_k^q)^2 = \mathrm{E}\,\left(x_k - (y_k^q + \widehat{x}_k)\right)^2 = \mathrm{E}\,\left(y_k + \widehat{x}_k - y_k^q - \widehat{x}_k\right)^2 = E(y_k - y_k^q)^2 = \mathrm{E}\,\widetilde{y}_k^2. \tag{2.16}$$

Noise feedback analysis

As mentionned in [14], analytical evaluation of noise feedback can be found in the literature [68, 69]. An important result is that the quantization noise feedback has very little effect on the optimal value of the prediction coefficient, *however coarse the quantization.* In the case of an AR(1) process, the optimal predictor $\bar{l}$ equals ρ. Assuming that the quantization noise $\widetilde{y}$ is white and decorrelated from the input of the quantizer, the prediction error variance in the closed loop is then[5]

$$\begin{aligned} {\sigma'}_y^2 = \mathrm{E}\,y_k^2 &= \mathrm{E}\,(x_k - \rho x_{k-1}^q)^2 \\ &= \mathrm{E}\,([x_k - \rho x_{k-1}] + \rho \widetilde{y}_{k-1})^2 \\ &\approx \sigma_x^2(1-\rho^2) + c2^{-2r}\rho^2 {\sigma'}_y^2 \\ &\approx \sigma_y^2 + c2^{-2r}\rho^2 {\sigma'}_y^2, \end{aligned} \tag{2.17}$$

where σ_y^2 is the optimal prediction error variance obtained by using unquantized data. This leads to

$${\sigma'}_y^2 \approx \frac{\sigma_y^2}{1 - c2^{-2r}\rho^2}, \tag{2.18}$$

which may be approximated as

$${\sigma'}_y^2 \approx \sigma_{y^0}^2(1 + c2^{-2r}\rho^2). \tag{2.19}$$

The quantization noise of an ideal coding scheme without feedback $\sigma_q^2 = c2^{-2r}\sigma_y^2$ is filtered by the energy of the optimal predictor, which increases the prediction error variance and decreases thereby the coding performance.

[5] As far as the variances are concerned, the susbscript $'$ will denote the presence of noise feedback in the rest of this chapter.

To summarize this analysis, describing the operational distortion-rate function of an entropy coded scalar quantizer by

$$d_x(r) = c2^{-2r}\sigma_x^2, \tag{2.20}$$

the distortion-rate function of an ideal DPCM coding scheme (neglecting noise feedback) would be

$$d_y(r) = c2^{-2r}\sigma_y^2. \tag{2.21}$$

The distortion-rate function of the DPCM coding scheme of figure 2.4 with noise feedback is evaluated by

$$d'_y(r) \approx c2^{-2r}{\sigma'}_y^2 \approx c2^{-2r}\frac{\sigma_y^2}{1-c2^{-2r}\rho^2}, \tag{2.22}$$

which from (2.19) and (2.21) may be further approximated as

$$d'_y(r) \approx d_y(r)\left(1+c2^{-2r}\rho^2\right) = d_y(r)\left(1+\frac{\rho^2 d_y(r)}{\sigma_y^2}\right). \tag{2.23}$$

Some Numerical Results

For entropy coded uniform scalar quantization (ECUQ) of a Gaussian source x, the Rényi's relation of differential to discrete entropy yields

$$r_i = H(x_i^q) \approx \frac{1}{2}\log_2 2\pi e\sigma_x^2 - \log_2\Delta. \tag{2.24}$$

Assuming sufficiently fine quantization with stepsize Δ, the distortion is $d_x \approx \frac{\Delta^2}{12}$[70], therefore

$$r_i \approx \frac{1}{2}\log_2\frac{2\pi e}{12 d_x} \Rightarrow d_x \approx \frac{\pi e}{6}2^{-2r_i}\sigma_{x_i}^2, \tag{2.25}$$

from which we see that c equals $\frac{\pi e}{6}$ for sufficiently high rate. From figure 2.5, which compares the actual performance coefficient c of an ECUQ to the high rate approximation, high rate means approximately 3 b/s. The distortion of ECUQ equals the Shannon lower bound at zero rate only, and is about twice this bound at approximately 1 b/s.

The comparison of the results given by the noise feedback analysis in DPCM, with the performance of an actual system using ECUQ is shown in fig. 2.6 and 2.7, for an AR(1) process ($\rho = 0.97$).
In figure 2.6,

- *(1)* "High-Rate $d_x(r)$ of ECUQ" is the Gish and Pierce approximation ($c = \frac{\pi e}{6}$) of ECUQ for the source x,
- *(2)* "Actual $d_x(r)$ of ECUQ" is the actual peformance of ECUQ for the source x,
- *(3)* "High rate $d_y(r)$ DPCM" is the Gish and Pierce approximation of expression (2.21),
- *(4)* "$d'_y(r)$ DPCM theor. feed." is the theoretic evaluation (2.22) of DPCM with quanitzation noise feedback,
- *(5)* "$d'_y(r)$ DPCM theor. feed." is the approximation (2.23) of the previous expression,

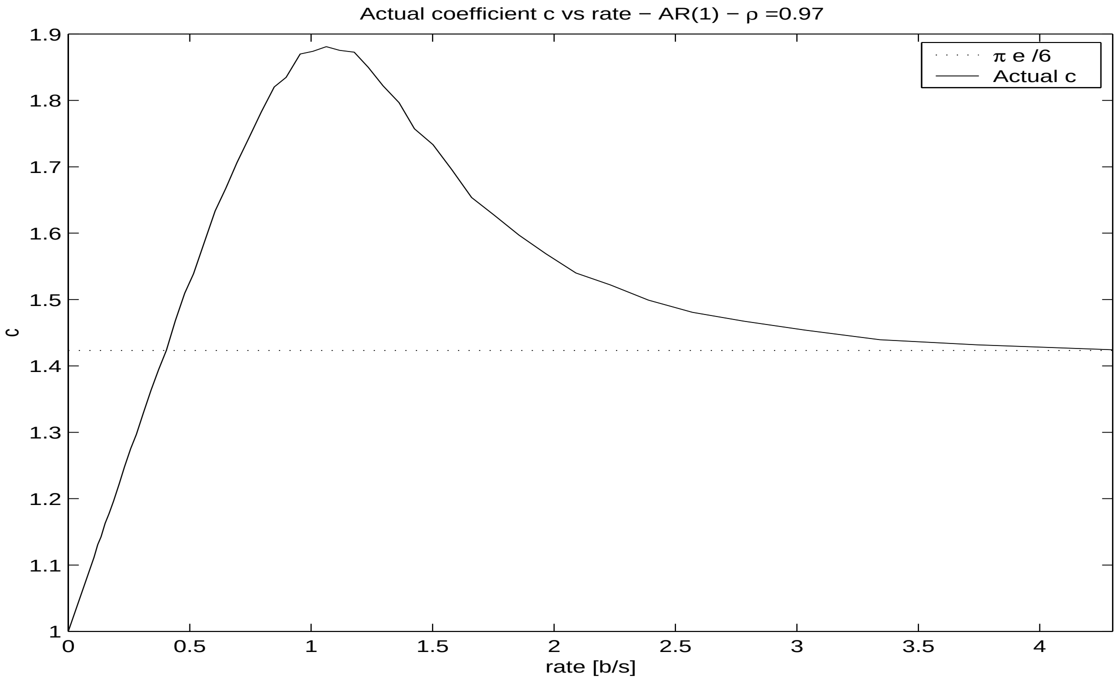


Figure 2.5: Comparison between actual value and high rate approximation of coefficient c for ECUQ.

- *(6)* " Actual $d'_y(r)$ DPCM ρ" is the actual performance of a DPCM system using predictor ρ and ECUQ of the prediction residual,
- *(7)* " Actual $d'_y(r)$ DPCM ρ'" is the actual performance of a DPCM system using a predictor optimized for the closed loop,
- *(8)* " Actual $d_y(r)$ without feed." is the actual rate distortion of the optimal prediction error signal (expression (2.14)),
- *(9)* " Actual $d_x(r)$ DPCM Open loop" is the actual distortion-rate function of the open loop system (expression 2.15).

As can be noted from the curves *(6)* and *(1)*, closed loop DPCM followed by entropy coding is, on the one hand, advantageous w.r.t. direct quantization and entropy coding even at low rates such as 0.5 b/s. On the other hand, an open loop system is useless (curve *(9)*).

Comparing *(6)* and *(8)* allows one to precisely observe which perturbation in the distortion of DPCM is caused by quantization noise (curve *(8)* may be seen as an hypothetic model of what would be the distortion if there were no feedback). Comparing with *(3)*, the effects of nonconstant c and of noise feedback become visible for r less than ≈ 3 b/s and increase fast when the rate decreases. These effects are well described by the first order perturbation analysis (curve *(5)*) for rates higher than approximately 1.5b/s, and even at lower rates by the more accurate exp. (curve *(4)*).

Finally, as previously observed in [68, 69], quantization feedback has very little effect on the optimal value of the prediction coefficient however coarse the quantization, and systems that use an optimized predictor

ρ' have almost the same performance (curve *(7)*) as a system designed with the optimal predictor without feedback ρ (curve *(6)*).

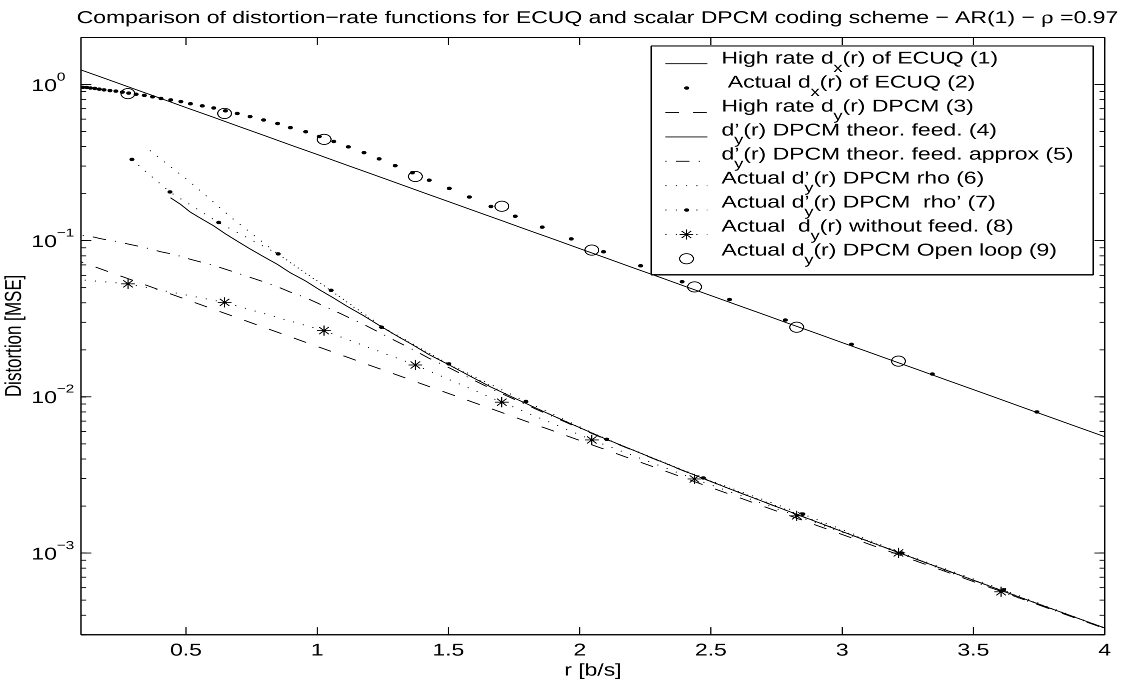


Figure 2.6: Comparison between theroetic and actual distortion-rate functions for Entropy Coded Uniform Quantization (ECUQ), and DPCM with ECUQ

In fig. 2.7, similar observations can be made from the curves corresponding to the SNR.

2.5.2 Quantization Noise Feedback in a Closed loop Causal Transform Coding Scheme

Let us now evaluate, for the causal transform, the perturbation caused by the quantization feeedback. In order to compute the expression of the coding gain in this case, the analysis of this section will be based on high resolution assumptions: all quantizers are assumed to have the same distortion-rate law $c2^{-2r_i}{\sigma'}^2_{y_i}$, with constant performance coefficient $c = \frac{\pi e}{6}$; ${\sigma'}^2_{y_i}$ are the actual prediction error variances obtained in the presence of noise feedback. Furthermore, we assume an optimal bit assignment. In the case where the reference vector is not based on the original signal but on its quantized version, the output vector becomes

$$\underline{y}_k = \underline{x}_k - \overline{L}\underline{x}^q_k = \underline{x}_k - \overline{L}(\underline{x}_k - \underline{\tilde{x}}_k) = L\underline{x}_k + \overline{L}\underline{\tilde{y}}_k\,. \tag{2.26}$$

The difference vector $\underline{y}_k$ now not only contains the prediction error $L\underline{x}_k$ of $\underline{x}_k$, but also the quantization error $\underline{\tilde{y}}_k$ filtered by the predictor $\overline{L}$. Equivalent representations of the closed loop causal coding schemes are

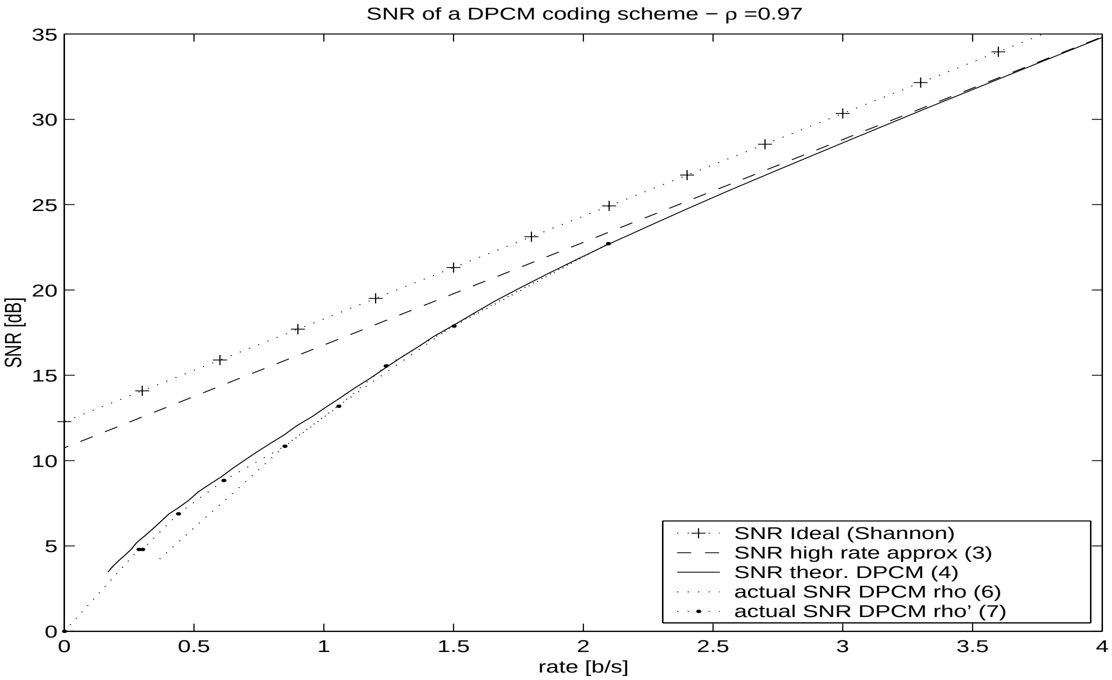


Figure 2.7: Comparison between actual and theoretic SNRs for DPCM

depicted in figures 2.8 and 2.9. In figure 2.9, the transform components are coded without reconstructing the data, that is, by using the quantized whitened versions y_i^q instead of x_i^q to compute the prediction.

As in the DPCM analysis, we will denote by the superscript $'$ the quantities relative to the case of noise feedback.

The optimization of the coding sheme is again comprised of two steps, optimal bit assignment and optimization the transform. This transform will be denoted by L' because it may be different from the transform L designed for a system without feedback. We will see however that as in DPCM, the optimal predictor should essentially vary.

As for the optimal bit assignment, it should again minimize the sum of the $\sigma'^2_{\tilde{y}_i}$. The variances of the quantization noises are therefore

$$\sigma'^2_{\tilde{y}_i} = c2^{-2r}(\prod_{i=1}^{N} \sigma'^2_{y_i})^{\frac{1}{N}} = \sigma'^2_q, \tag{2.27}$$

and the autocorrelation matrix of the noise is $R'_{\underline{\tilde{y}\tilde{y}}} = \sigma'^2_q I$. Comparing with (2.6), the variances of the transform signals are increased because the reference vector is based on quantized data, and the quantization noise components are therefore increased to σ'^2_q.

We shall now optimize L'. The coding gain for an optimal bit allocation is then, as usual,

$$G^{(1)}_{L'} = \frac{\mathrm{E}\,||\tilde{x}_k||^2}{\mathrm{E}\,||\tilde{y}_k||'^2_{L'}} = \left(\frac{\det\,\mathrm{diag}\,\{R_{\underline{xx}}\}}{\det\,\mathrm{diag}\,\{R'_{\underline{yy}}\}}\right)^{\frac{1}{N}}, \tag{2.28}$$

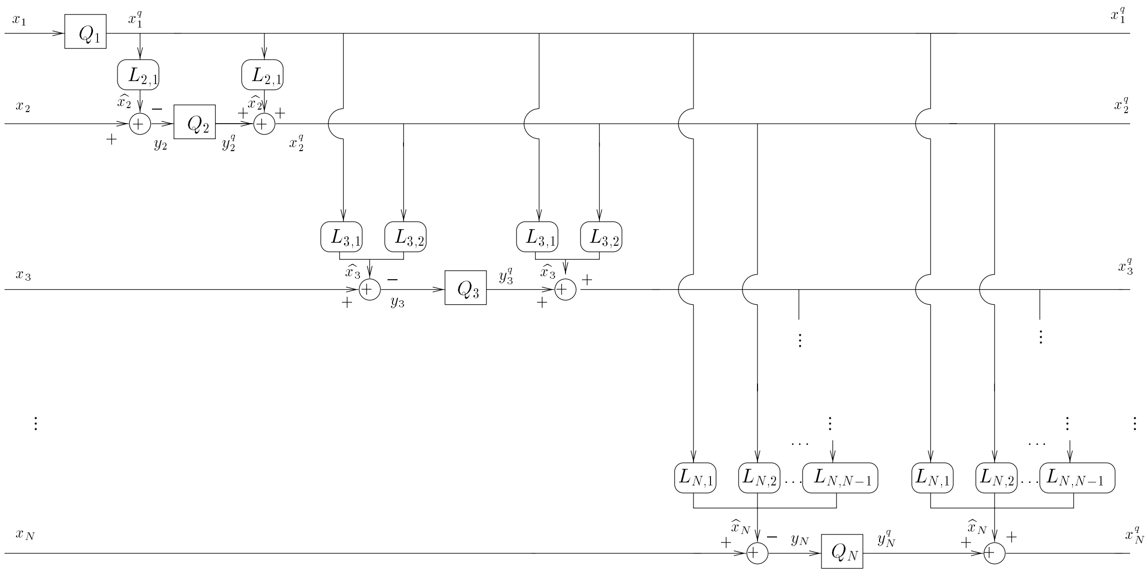


Figure 2.8: Closed loop causal transform coding sheme.

where notation $^{(1)}$, as opposed to $^{(0)}$ in (2.10), refers to noise feedback. The variances of the transform signals $(R'_{\underline{yy}})_{ii} = \sigma'^2_{y_i}$ depend now on L', and in order to optimize L', one should consider

$$\min_{L'} \left(\det\left[\text{diag}\,\{R'_{\underline{yy}}\}\right]\right) \ . \tag{2.29}$$

This time, $\underline{y}_k = \underline{x}_k - L'\underline{x}_k^q = L'\underline{x}_k - \overline{L}'\widetilde{x}_k$, and

$$R'_{\underline{yy}} = L'R_{\underline{xx}}L'^T + \sigma'^2_q\overline{L}'\overline{L}'^T \ . \tag{2.30}$$

Since $L' = I - \overline{L}'$, $\overline{L}'\overline{L}'^T = L'L'^T - I + \overline{L}' + \overline{L}'^T$, and since the prediction matrix $\overline{L}'$ is strictly lower triangular, we get

$$\text{diag}\,\{\sigma'^2_q\overline{L}'\overline{L}'^T\} = \text{diag}\,\{L(\sigma'^2_q I)\,L^T - \sigma'^2_q I\}, \tag{2.31}$$

and it follows that

$$\det\left[\text{diag}\,\{R'_{\underline{yy}}\}\right] = \det\left[\text{diag}\,\{L'(R_{\underline{xx}} + \sigma'^2_q I)\,L'^T - \sigma'^2_q I\}\right] \tag{2.32}$$

This problem is again separable and corresponds, for the purpose of this analysis taking into account the first order of the perturbation, to the optimal prediction of $\underline{x}$ perturbated by a white noise. Note that this does not imply that $R_{\underline{x}^q\underline{x}^q} = R_{\underline{xx}} + \sigma'^2_q I$ (this would be the case for high resolution ECUQ, but not for optimal FRQ, in which case the variance of the quantized signals are less than the original ones).

In order to optimizing L', we should look for

$$\min_{L'_{i,1:i-1}} L'_i(R_{\underline{xx}} + \sigma'^2_q I)L'^T_i \ . \tag{2.33}$$

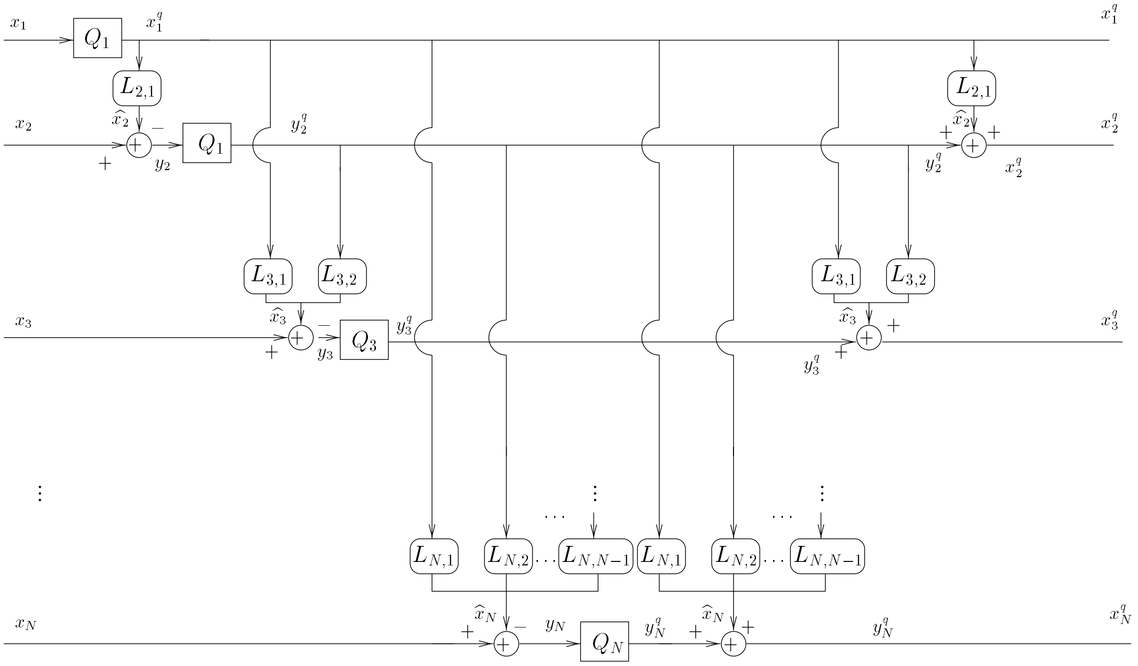


Figure 2.9: Closed loop causal transform based on whitened quantized data.

For this problem, one can denote the resulting optimal prediction error variances ${\sigma'}^2_q + {\sigma'}^2_{y_i}$ where ${\sigma'}^2_{y_i}$ is the variances required for the coding gain. These variances may further be written as ${\sigma'}^2_{y_i} = \sigma^2_{y_i} + \Delta\sigma^2_{y_i}$, where $\sigma^2_{y_i}$ is obviously the optimal prediction error variance without quantization noise feedback, and $\Delta\sigma^2_{y_i}$ denotes the contribution to the prediction error variance of the quantization noise on the previous samples . The normal equations for (2.33) can hence be written as[6]

$$\left[\begin{array}{c} \\ R_{1:i,1:i} + {\sigma'}^2_q I_i \\ \\ \\ \end{array}\right] \left[\begin{array}{c} L'_{i,i-1} \\ \vdots \\ L'_{i,1} \\ 1 \end{array}\right] = \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ {\sigma'}^2_q + {\sigma'}^2_{y_i} \end{array}\right],$$

wich leads to

$${\sigma'}^2_{y_i} + {\sigma'}^2_q = R_{i,i} + {\sigma'}^2_q - R_{i,1:i-1}(R_{1:i-1,1:i-1} + {\sigma'}^2_q I_i)^{-1} R_{1:i-1,1}. \tag{2.34}$$

Under the high resolution assumption, the term ${\sigma'}^2_q I$ is small in comparison with $R_{1:i-1,1:i-1}$, and we get the approximation up to first order of perturbation

$${\sigma'}^2_{y_i} \approx \underbrace{R_{i,i} - R_{i,1:i-1} R^{-1}_{1:i-1,1:i-1} R_{1:i-1,i}}_{=\sigma^2_{y_i}} + \underbrace{{\sigma'}^2_q \underbrace{||R^{-1}_{1:i-1,1:i-1} R_{1:i-1,1}||^2}_{=||L_{i,1:i-1}||^2 = ||L_i||^2 - 1 = (\overline{LL}^T)_{ii}}}_{\Delta\sigma^2_{y_i}}, \tag{2.35}$$

[6] For notation simplicity R denotes $R_{\underline{xx}}$ until eq. (2.35).

where $\sigma^2_{y_i}$ and $\overline{L}$ are non perturbed quantities.
Hence, we get

$$\begin{aligned} {\sigma'}^2_{y_i} &\approx (LR_{\underline{xx}}L^T + {\sigma'}^2_q\overline{LL}^T)_{ii} \\ &\approx (LR_{\underline{xx}}L^T + \sigma^2_q\overline{LL}^T)_{ii} \end{aligned} \tag{2.36}$$

where L and $\overline{L}$ are non perturbed quantities.
Suppose now that the transform L of section (2.3), optimized without feedback, is used to compute the reference vectors in a closed loop coding scheme. Then the variance of the transform signals will also be given by (2.36). This suggests that the optimal predictor design should not essentially vary, at least at moderate to high rates.
Summarizing, the distortion (2.27) is

$$\begin{aligned} \tfrac{1}{N}\,\mathrm{E}\,||\underline{\tilde{y}}||'^2_{L'} &\approx \tfrac{1}{N}\,\mathrm{E}\,||\underline{\tilde{y}}||'^2_{L} \\ &\approx \tfrac{1}{N}\sum_{i=1}^{N}\frac{\pi e}{6}2^{-2r_i}{\sigma'}^2_{y_i} \\ &\approx \tfrac{\pi e}{6}2^{-2r}\left(\prod_{i=1}^{N}{\sigma'}^2_{y_i}\right)^{\frac{1}{N}} . \end{aligned} \tag{2.37}$$

Using (2.36), this equation may be rewritten as

$$\begin{aligned} \tfrac{1}{N}\,\mathrm{E}\,||\underline{\tilde{y}}||'^2_{L',L} &\approx \tfrac{\pi e}{6}2^{-2r}\left(\det\ \text{diag}\,\{LR_{\underline{xx}}L^T + \sigma^2_q\overline{LL}^T\}\right)^{\frac{1}{N}} \\ &\approx \tfrac{\pi e}{6}2^{-2r}\left(\prod_{i=1}^{N}\sigma^2_{y_i}(1+\sigma^2_q\frac{||\overline{L}_i||^2}{\sigma^2_{y_i}})\right)^{\frac{1}{N}} \\ &\approx \sigma^2_q\left(1+\tfrac{\sigma^2_q}{N}\sum_{i=1}^{N}\frac{||\overline{L}_i||^2}{\sigma^2_{y_i}}\right) . \end{aligned} \tag{2.38}$$

Referring to (2.28), this leads to the following expression for the coding gain $G^{(1)}_L$, taking into account the perturbations up to first order

$$G^{(1)}_L \approx G^{(1)}_{L'} \approx \left(\frac{\det\left[\,\text{diag}\,\{R_{\underline{xx}}\}\right]}{\det\left[\,\text{diag}\,\{LR_{\underline{xx}}L^T + \sigma^2_q\overline{LL}^T\}\right]}\right)^{\frac{1}{N}} \approx G^{(0)}_L\left(1-\frac{1}{N}\sigma^2_q\sum_{i=1}^{N}\frac{||\overline{L}_i||^2}{\sigma^2_{y_i}}\right) . \tag{2.39}$$

with $LR_{\underline{xx}}L^T = R_{\underline{yy}}$, $\sigma^2_q = c\,2^{-2R}(\det R_{\underline{xx}})^{\frac{1}{N}}$, $R_{\underline{yy}}$ is the diagonal matrix of the non perturbated prediction error variances, and L and $\overline{L}$ are also non perturbated quantities. This expression is established under high resolution assumptions (small quantization noise variance w.r.t. signal variances, and performance factor c is constant).
An equivalent and interesting expression of $G^{(1)}_L$ is (see Sec. 2.C)

$$G^{(1)}_L \approx G^{(0)}_L\left(1-\frac{\sigma^2_q}{N}\sum_{i=1}^{N}\left[\frac{1}{\lambda_i}-\frac{1}{\sigma^2_{y_i}}\right]\right) \tag{2.40}$$

where $G^{(0)}L$ is the coding gain in the ideal case, $\{\lambda_i\}$ are the eigenvalues of the autocorrelation matrix of $\underline{x}$, and σ^2_q is the quantization noise in the ideal case, assumed to be white. Thus, maximizing the coding gain entails maximizing the sum of the inverses of the prediction error variances. Whereas the coding gain in the ideal case is invariant by permutation, there is in the closed loop causal transform coding scheme an optimal ordering of the components of the $\underline{x}_i$. It will be shown in chapter 5 that maximizing $G^{(1)}$ entails decorrelating the signals x_i by order of decreasing variances.

2.6 Analysis of a Practical Case

A simple mean of realizing nearly optimal bit assignment in the case of ECUQ is to quantize the signals with equal quantization stepsizes. This case allows one to check, for a practical transform coding system, firstly in which range of average rates the LDU suffers from a non negligible noise feedback, and for which rates it presents similar coding performances to those of the KLT; secondly if the previously exposed noise analysis has some value in this practical case; thirdly if the claimed, and not proven yet, decorrelation strategy (consisting in processing the signals by order of decreasing variances) is actually the best one.

2.6.1 Optimal Bit Assignment and Equal Quantization Stepsize

The classical high rate result of the optimal bit assignment states that, given a set of variances $\sigma^2_{y_i}$, the quantization noise $\sigma^2_{q_i}$ should be equal for all the components. The number of bits assigned to the ith component is then

$$r_i = r + \frac{1}{2}\log_2 \frac{\sigma^2_{y_i}}{\left(\prod_{i=1}^{N}\sigma^2_{y_i}\right)^{\frac{1}{N}}}. \tag{2.41}$$

Under high resolution assumption, the quantization noise resulting from quantization with stepsize Δ_i is uniformly distributed with variance $\sigma^2_{q_i} = \frac{\Delta_i^2}{12}$. A simple way of realizing equal distortion is therefore to quantize all the components with an equal stepsize Δ. If the y_i^q are further entropy coded, the bitrate r_i is given by

$$r_i = H(y_i^q) \approx \frac{1}{2}\log_2 2\pi e\sigma^2_{y_i} - \log_2\Delta. \tag{2.42}$$

It can then easily be checked that choosing

$$\Delta = \sqrt{2\pi e}2^{-r}(\prod_{i=1}^{N}\sigma^2_{y_i})^{\frac{1}{2N}} = \sqrt{2\pi e}2^{-r}(\prod_{i=1}^{N}\sigma^2_{y_i})^{\frac{1}{2N}} \tag{2.43}$$

corresponds to $\frac{1}{N}\sum_{i=1}^{N} r_i = \frac{1}{N}\sum_{i=1}^{N} H(y_i^q) \approx r$. At high rates, the corresponding distortion-rate function is then $\frac{\Delta^2}{12} \approx \frac{\pi e}{6}2^{-2r}(\prod_{i=1}^{N}\sigma^2_{y_i})^{\frac{1}{N}}$. For a particular two-dimensional Gaussian source obtained by means of a KLT, numerical simulations showed that optimal bit allocation and equal quantization stepsizes are equivalent as long as the rate dedicated to each component is at least 1 bit per sample[7] [54].

[7] The corresponding average rate may be much higher.

2.6.2 Distortion Analysis

For large quantization stepsizes (low rates), the relation of differential to discrete entropy as given by (2.42) may not be accurate[8], be there a noise feedback or not. Thus, in the case of transform signals obtained in a closed loop system working at moderate to low rates, the average distortion

$$\frac{1}{N}\,\mathrm{E}\,||\underline{\widetilde{y}}||'^2 = \frac{1}{N}\sum_{i=1}^{N} d'_{i,L} = \frac{1}{N}\sum_{i=1}^{N} c2^{-2r_i}\sigma'^2_{y_i} \tag{2.44}$$

may be different from the geometric mean (2.37). This analysis will be guided again by that of DPCM systems. In order to describe the coding system implemented in closed loop, we will refer to perturbation w.r.t. an open loop system. For this system, the distortion of each component is $\mathrm{E}\,\widetilde{y}_i^2 = d_{i,L} = c2^{-2r_i}\sigma^2_{y_i}$, where $\sigma^2_{y_i}$ are the variances of the optimal prediction errors. The average distortion for the open loop system in the transform domain is then

$$\frac{1}{N}\,\mathrm{E}\,||\underline{\widetilde{y}}_i||^2 = \frac{1}{N}\sum_{i=1}^{N} {d_{i,L}}^2 = \frac{1}{N}\sum_{i=1}^{N} c2^{-2r_i}\sigma^2_{y_i}, \tag{2.45}$$

Figure 2.10 compares the operational distortion-rate function (2.45) obtained with equal quantization stepsize for signals of decreasing and increasing variances. The high rate and optimal bit allocation approximation $\frac{\pi e}{6}2^{-2r}(\det R_{\underline{xx}})^{\frac{1}{N}}$ is plotted in full line. It can be seen that even without noise feedback, the average distortion (2.45) deviates noticeably from the high rate and optimal bit allocation approximations at rates lower than approximately 3 b/s.

For a closed loop causal transform system now, we assume that the covariance matrix of the quantization noise is well approximated by $\frac{\Delta^2}{12}I$ at moderate to high rates. Thus, the optimal transform is again given by (2.33). From (2.35), the actual prediction error variances σ'_{y_i} may still be approximated by expression (2.36. Again, similar performance should be obtained for small perturbations by using either the optimal transform L' (minimizing (2.33)), or the transform designed without feedback L. Using (2.36), the operational distortion-rate function of the transform signals with quantization noise feedback may be evaluated as

$$\begin{aligned} d'_{i,L} &\approx \tfrac{\Delta^2}{12} \\ &\approx c2^{-2r_i}\sigma'^2_{y_i} \\ &\approx c2^{-2r_i}(\sigma^2_{y_i} + \underbrace{\frac{\Delta^2}{12}}_{d'_{i,L}}(\overline{LL^T})_{ii})) \\ d'_{i,L}(1 - ||\overline{L_i}||^2 c2^{-2r_i}) &\approx d_{i,L} \\ d'_{i,L} &\approx \frac{d'_{i,L}}{1 - ||\overline{L_i}||^2 c2^{-2r_i}} \end{aligned} \tag{2.46}$$

[8] The discrepancy with the actual distortion is apparent at zero rate in (2.25).

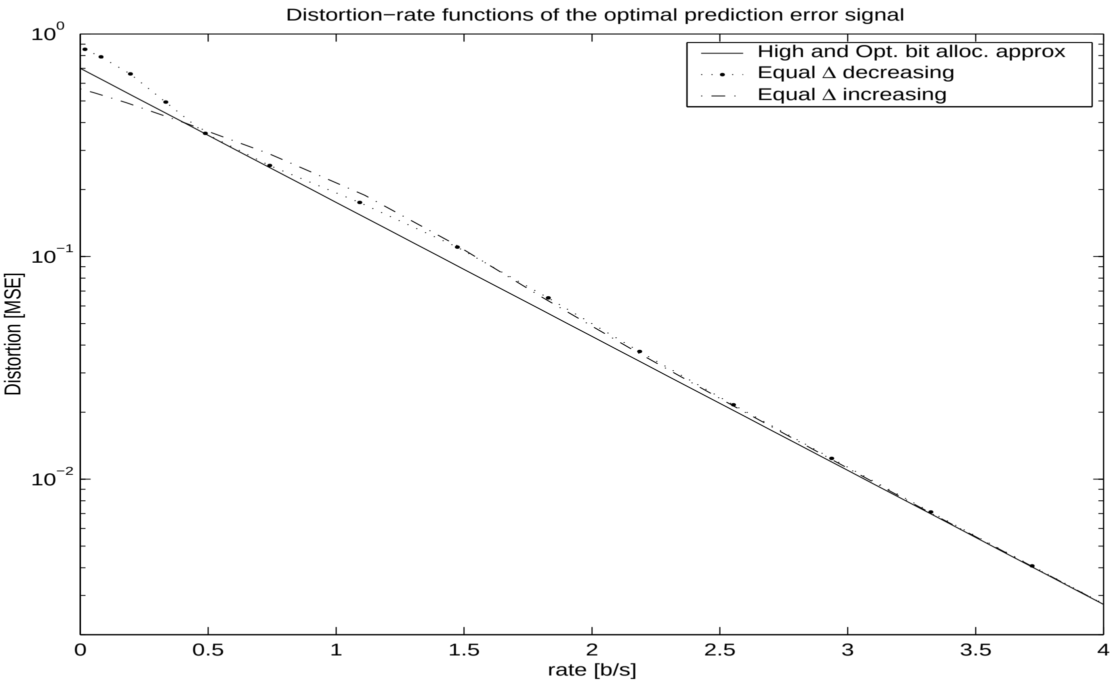


Figure 2.10: Distortion-rate functions of the optimal prediction error signals.

Hence,

$$\begin{aligned}\tfrac{1}{N}\,\mathrm{E}\,\|\underline{\tilde{y}}\|'^{2}_{L} &= \tfrac{1}{N}\sum_{i=1}^{N} c2^{-2r_i}{\sigma'}^{2}_{y_i} \\ &\approx \tfrac{1}{N}\sum_{i=1}^{N}\frac{d_{i,L}}{1-\|\overline{L_i}\|^2 c2^{-2r_i}} \\ &\approx \tfrac{1}{N}\sum_{i=1}^{N}\frac{d_{i,L}}{1-\|\overline{L_i}\|^2\frac{d_{i,L}}{\sigma^2_{y_i}}} \\ &\approx \tfrac{1}{N}\sum_{i=1}^{N} d_{i,L}(1+\|\overline{L_i}\|^2\frac{d_{i,L}}{\sigma^2_{y_i}}).\end{aligned} \tag{2.47}$$

At high rates, the distortions $d_{i,L}$ tend to $\frac{\pi e}{6}2^{-2r_i}\sigma^2_{y_i} = \frac{\pi e}{6}2^{-2r}(\det R_{\underline{xx}})^{\frac{1}{N}} = \sigma^2_q$, and the above distortion tends to (2.38).

2.6.3 Numerical results

The data are real Gaussian i.i.d. vectors with covariance matrix $R = HR_{AR1}H^T$. R_{AR1} is the covariance matrix of an AR(1) process with parameter $\rho = 0.9$. H is a diagonal matrix whose ith entry is $(N-i+1)^{1/3}$, $N = 3$. The signals x_i are coded by order of either decreasing, or increasing variances. For these two decorrelating strategies, sets of 10^4 vectors were transformed using each of the two causal closed loop algorithms depicted in fig. (2.8) (on the basis of reconstructed data) and (2.9) (quantized whitened data)[9]. The resulting transform signals where then quantized using the same stepsize Δ. For each component, we measured the entropies of the discrete valued signals y_i^q. The reconstructed vectors where then used to compute the average distortion over the whole data set (of length 10^4). For a given Δ, we repeated this experiment ten times. Several values of Δ where investigated in order to cover a wide range of rates.

- In fig. 2.11 and 2.12, the distortion-rate functions of the closed loop causal transform are plotted for signals of decreasing and increasing variances respectively (optimal transform L of eq. (2.8) is used)

  - *(1)* “Theoretic with feedback” refers to the analytical evaluation (2.47) of a system with equal stepsize,
  - *(2)* “Actual with feedback” corresponds to the actual distortion-rate function of the resulting closed loop TC system,
  - *(3)* “Theoretic with Equal c and feedback” refers to the analytical evaluation (2.38) obtained for an optimal bit assignment algorithm,
  - *(4)* “High & Opt. bit alloc. approx” refers to the performance of a system without feedback, constant quantizer performance factor $c = \frac{\pi e}{6}$ and optimal bit allocation, as given in (2.9).

  It can be observed for both decorrelation strategies that

  - The performance of actual systems *(2)* deviate from their high rate approximation *(4)* for rates below approximately 3 b/s.
  - These performance are accurately described by the analysis (curve *(1)*) down to approximately 1 b/s.
  - The analysis of section 2.5.2, which does not account for possible variations of c w.r.t. the rate underestimates the actual distortions (it was shown in fig. 2.10 that even without noise feedback, the actual distortion in the transform domain with equal Δ is larger than (2.9)).

  Comparing now the figures 2.11 and 2.12, it is clear that better performance are obtained by processing the signals by order of decreasing variance, as suggested by the high rate analysis of section 2.5.2.

- In fig. 2.13 and 2.14, the actual distortion-rate performance for algorithms based either on reconstructed, or on quantized and whitened data are compared for decreasing, and increasing variances respectively. It can be seen that the two approaches yield essentially the same performance.

- In fig. 2.15 and 2.16, the actual distortion-rate performance for algorithms using either the causal transform optimized without feedback (2.8), or the predictor L' optimized as in (2.33) are compared

[9] Unless otherwise stated, the presented results are based on the algorithm using reconstructed data. The performances of both algorithms are similar, see fig. 2.13 and 2.14

for decreasing, and increasing variances respectively. It can be observed that even at low rates, the two transforms yield comparable performance, however coarse the quantization. This result is similar to that reported for DPCM [14].

- Finally, fig. 2.17 presents a comparison of systems using either no transform, or the KLT, or the LDU:
  - *(1)* “Actual D without transform” refers to directly entropy coding the x_i,
  - *(2)* “High rate Approx D without transform” refers to the high rate (constant c and optimal bit assignment) approximation of *(1)*,
  - *(3)* “Actual D LDU decreasing” refers to the actual distortion of the causal closed loop transform coding scheme processing the signals by order of decreasing variances (predictor L is used),
  - *(4)* “Actual D LDU increasing” as in *(3)* but with increasing variances (predictor L is used),
  - *(5)* “ Actual D KLT increasing/decreasing” refers to the actual distortion of a system using the KLT (since there is no noise feedback, the distortion is invariant by permutation),
  - *(6)* “High rate and opt. bit alloc approx” refers to the performance of a system without feedback, and with constant quantizer performance factor (2.9).

  Note that the performance of the LDU is inferior to that of the KLT at low rates only (approximately 2 b/s). The LDU with either a decreasing- or increasing-variance decorrelation strategy is still advantageous (w.r.t. direct entropy coding the signals) at all rates.

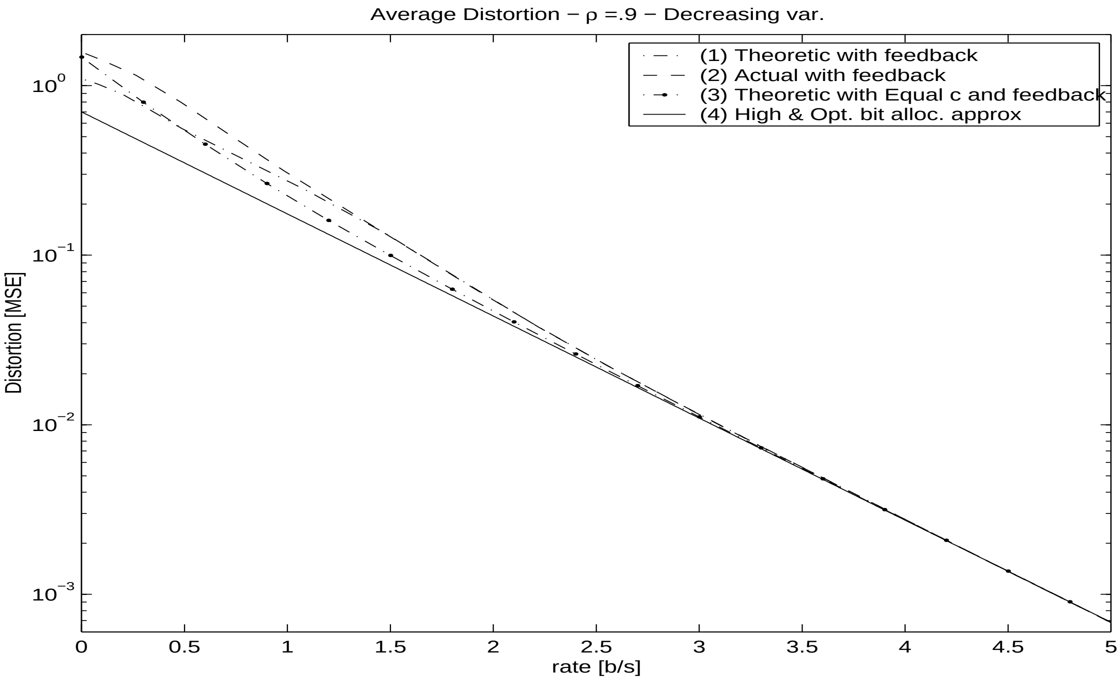


Figure 2.11: Average distortion *vs* rate for the causal transform with equal quantization stepsize (decreasing variances).

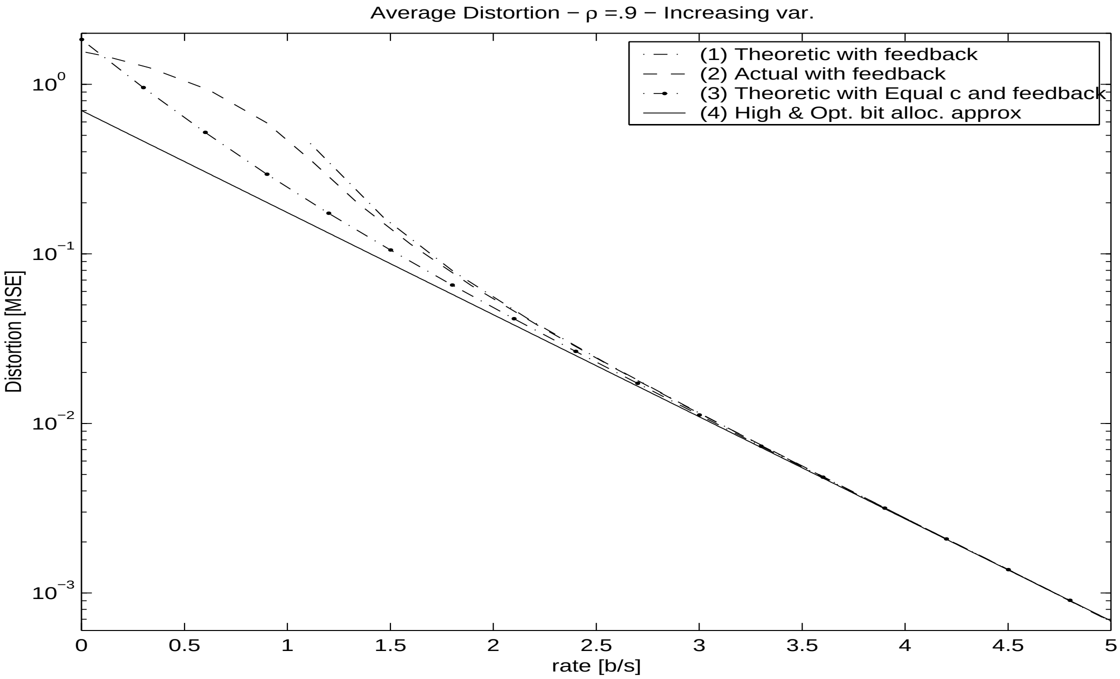


Figure 2.12: Average distortion *vs* rate for the causal transform with equal quantization stepsize (increasing variances).

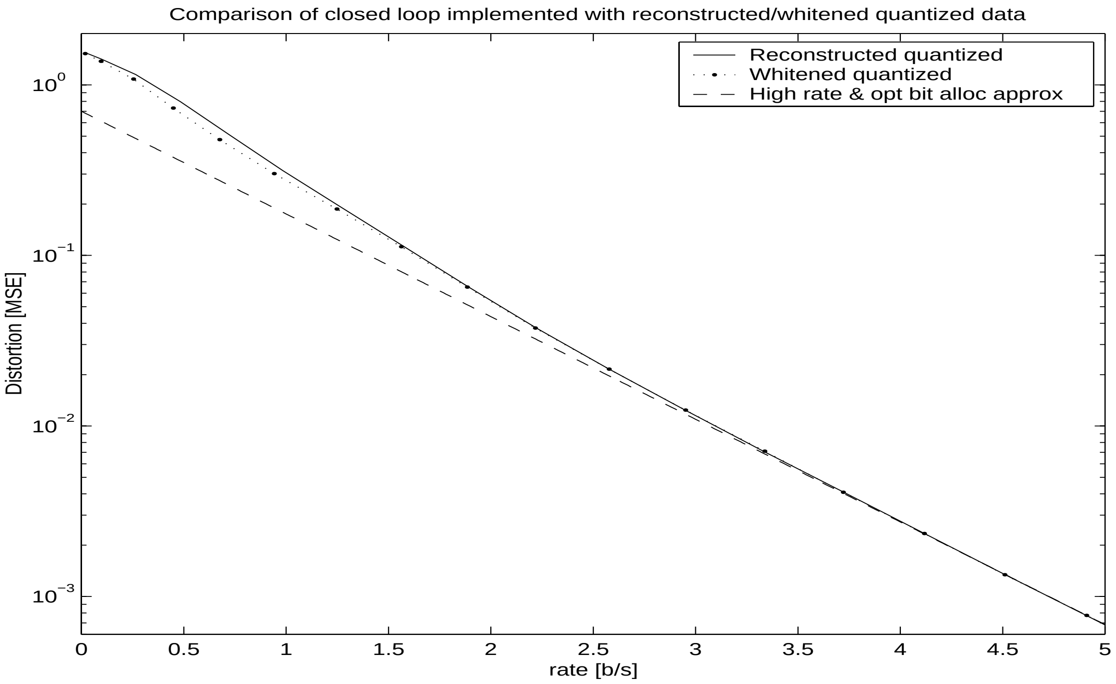


Figure 2.13: Comparison of actual distortion-rate performance for algorithms based either on reconstructed, or on quantized and whitened data (decreasing variances).

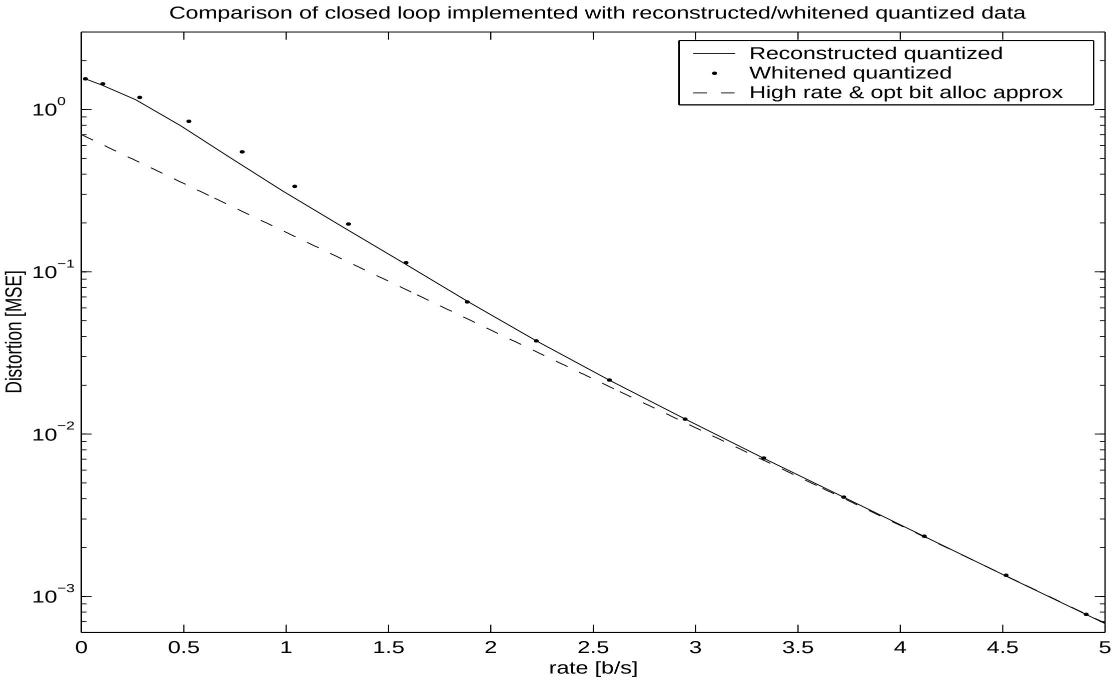


Figure 2.14: Comparison of actual distortion-rate performance for algorithms based either on reconstructed, or on quantized and whitened data (decreasing variances).

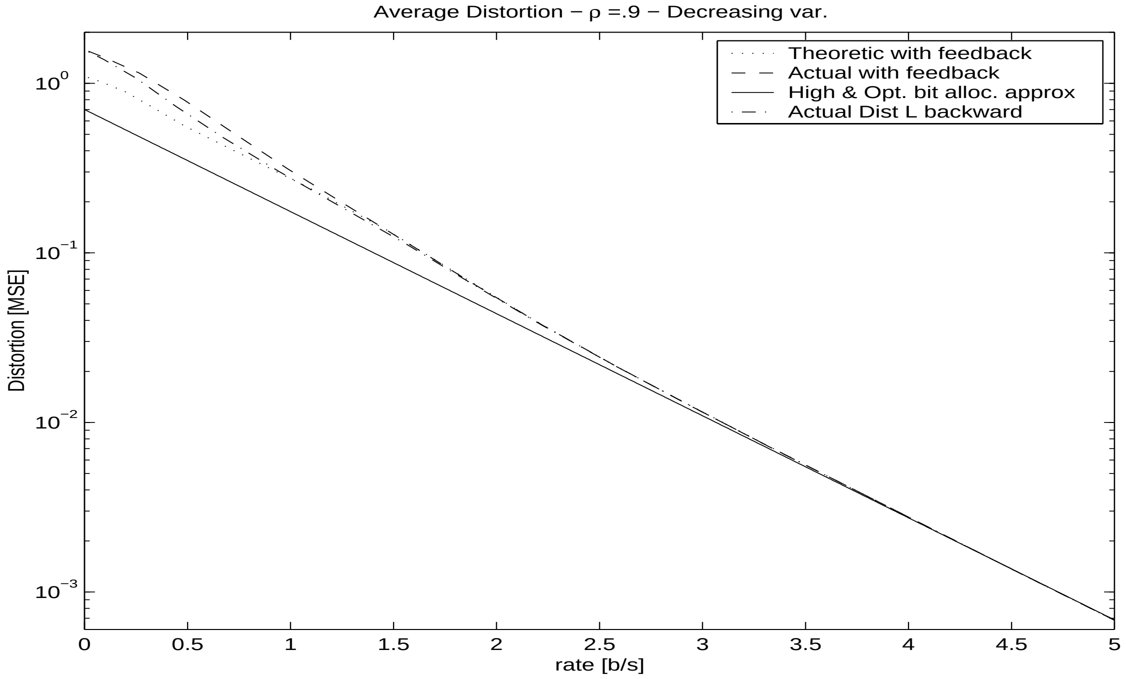


Figure 2.15: Comparison of actual distortion-rate performance for algorithms using either the transform L or L' (decreasing variances).

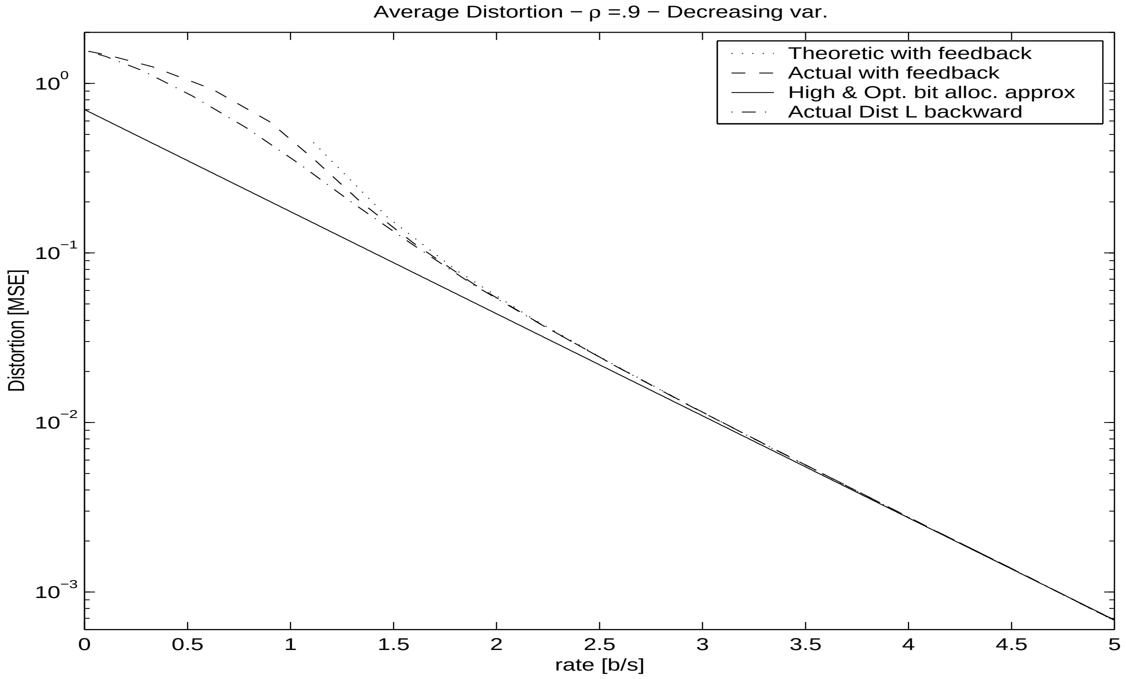


Figure 2.16: Comparison of actual distortion-rate performance for algorithms using either the transform L or L' (increasing variances).

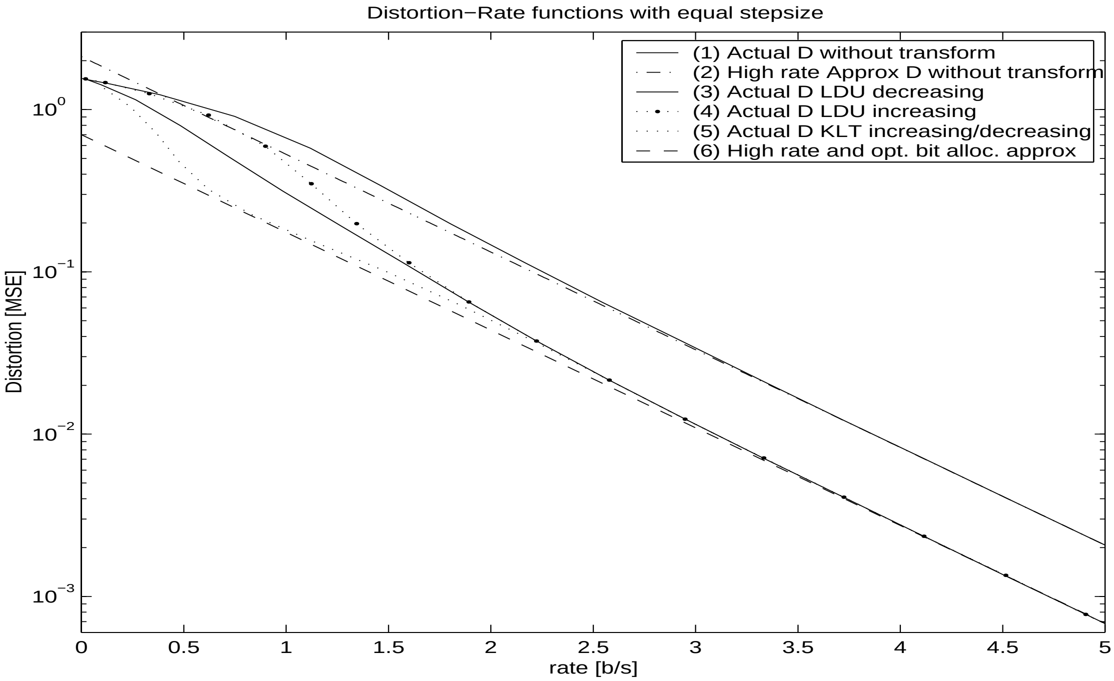


Figure 2.17: Comparison of the average distortions *vs* rate for the causal and the KL transforms with equal quantization stepsize(increasing and decreasing variances)

2.7 Conclusions

The optimal decorrelating transform subject to the constraint of causality was shown to correspond to an LDU factorization of the signal covariance matrix $R_{\underline{xx}}$. This transform was compared to the KLT and shown to provide asymptocically (w.r.t. the rate) the same coding gain as its unitary counterpart. Besides the equivalence of the coding gain at high rates, the optimal causal transform presents several advantages w.r.t. KLT, such as lower implementation and design complexities, and a best behaviour w.r.t. the quantization of the transform coefficients.

As in classical (A)DPCM, closed loop implementation of the causal coding structure was shown to be preferable. A high resolution analysis of the noise feedback effect uppon the coding gain was proposed in a first step. This analysis models these perturbation effects, assuming that they are small, and that the bits are optimally allocated. In a second step, an analytical evaluation of a practical transform coding algorithm was proposed. This transform coding scheme uses equal quantization stepsize, and entropy coded uniform quantizers. For this algorithm, the deviation from high rate approximation are noticeable beyond approximately 3 b/s for both the KLT and the LDU. In the causal case, the effects of the noise feedback become non negligible beyond approximately 2 b/s, and are well described by the proposed analysis. Moreover, decorrelating the signals by order of decreasing variances was shown empirically to be the best strategy. Comparing finally the two approaches, the LDU is shown to compete with the KLT at rates higher than approximately 2 b/s.

2.A Quantization noise model

The aim of this appendix is to provide some definitions and known results regarding unbounded uniform quantization of Gaussian sources. In particular, section 2.A.2 shows that the additive noise model may accurately describe the effects of quantization is this case, assuming sufficiently high rates.

2.A.1 Uniform Quantization

A quantizer can be viewed as a nonlinear mapping from the domain of continuous-amplitude inputs onto one of a possible outputs levels. The analysis of errors introduced by this mapping can be approached using stochastic methods. In this framework, the output of the quantizer is modeled as an infinite precision input and additive noise. The additive noise is a random variable whose distribution is nonzero only over an interval equal to the quantization stepsize. Widrow [71] showed that under the condition that the input r.v. has a band-limited characteristic function, the quantization is uniformly distributed; this is frequently refered to as the *quantization theorem*. For Gaussian inputs such as those considered in this work, this band-limitedness is not verified; theoretic results exist however, which precisely describes the statistical properties of the quantization noise and that of the reconstructed input [70, 72].

A *roundoff quantizer* of uniform step- (or *grain*-) size Δ has a staircase input-output relation (see fig. 2.18).

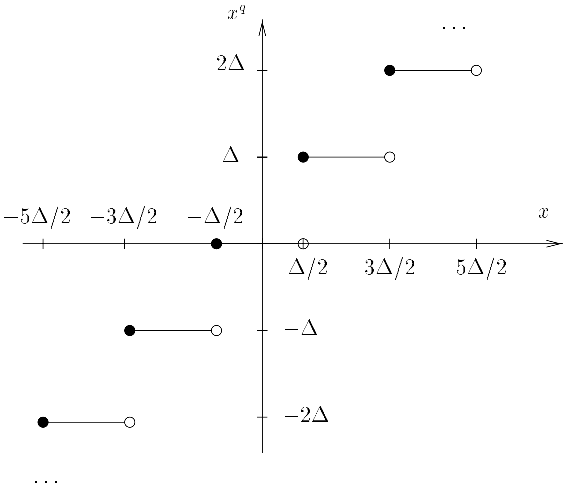


Figure 2.18: Input-output characteristic of an unbounded roundoff uniform quantizer.

At each sampling instant k, the input x_k, the quantized output x_k^q, and the quantization error $\widetilde{x}_k$ are related by

$$\widetilde{x}_k = x_k - x_k^q, \tag{2.48}$$

where $x_k^q = n\Delta$, $n \in \mathbb{Z}$.[10]

If we assume the input x to be a continuous variable with p.d.f. $f_x(x_k)$ and characteristic function

$$\phi_x(u) = \mathrm{E}\,(e^{jux}), \tag{2.49}$$

the p.d.f. of the quantization error $f_{\tilde{x}}$ is [70]

$$f_{\tilde{x}}(\tilde{x}_k) = \frac{1}{\Delta} + \frac{1}{\Delta}\sum_{n\neq 0}\phi_x\left(\frac{2\pi n}{\Delta}\right)e^{\frac{-2j\pi n\tilde{x}_k}{\Delta}} \quad \text{if} \quad -\frac{\Delta}{2} \leq \tilde{x}_k < \frac{\Delta}{2}, \text{ and } 0 \text{ otherwise.} \tag{2.50}$$

If the characteristic function satisfies

$$\phi_x(2\pi n/\Delta) = 0 \;\; \forall\; n \neq 0, \tag{2.51}$$

the p.d.f. is uniform (that is, equals $\frac{1}{\Delta}$ in the nontrivial interval). In this case, $E\tilde{x} = 0$, and $\mathrm{E}\,\tilde{x}^2 = \frac{\Delta^2}{12}$.

2.A.2 Gaussian sources

For zero mean Gaussian signals, the condition (2.51) is not satisfied. In this case, the p.d.f. is of the form

$$f_x(x_k) = \frac{1}{\sqrt{2\pi e}}e^{-\frac{x_k^2}{2\sigma_x^2}}, \tag{2.52}$$

where σ_x^2 is the variance of the input, the characteristic function is

$$\phi_x(u) = \mathrm{E}\,(e^{\frac{u^2\sigma_x^2}{2}}), \tag{2.53}$$

and the p.d.f of the error is given by

$$f_{\tilde{x}}(\tilde{x}_k) = \frac{1}{\Delta}\left[1 + 2\sum_{n=1}^{\infty}\cos\left(\frac{2\pi n\tilde{x}_k}{\Delta}\right)e^{-\frac{2\pi n^2\sigma_x^2}{\Delta^2}}\right] \quad \text{if} \quad -\frac{\Delta}{2} \leq \tilde{x}_k < \frac{\Delta}{2}, \text{ and } 0 \text{ otherwise.} \tag{2.54}$$

The mean of the quantization error is still zero, but the variance becomes

$$\mathrm{E}\,\tilde{x}^2 = \frac{\Delta^2}{12}\left[1 + \frac{12}{\pi^2}\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}e^{-\frac{2\pi n^2\sigma_x^2}{\Delta^2}}\right]. \tag{2.55}$$

Figure 2.19 plots the actual variance (2.55) and the $\frac{\Delta^2}{12}$ approximation. It can be seen that even for coarse quantization ($\Delta \approx 2\sigma$), the latter approximation is accurate. Figure 2.20 compares the actual correspondence between rate and Δ ("Rényi's correspondence" refers to that obtained assuming $H(x^q = \frac{1}{2}\log_2 2\pi e\sigma_x^2 - \log_2\Delta)$), the $\frac{\Delta^2}{12}$ approximation is accurate at rates as low as approximately 1 b/s.

The variance of the quantized r.v. x^q is given by

$$\mathrm{E}\,(x^q)^2 = \sigma_x^2 + 4\sigma_x^2\sum_{k=1}^{\infty}(-1)^k e^{-\frac{2\pi k^2\sigma_x^2}{\Delta^2}} + \frac{\Delta^2}{12}\left[1 + \frac{12}{\pi^2}\sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}e^{-\frac{2\pi k^2\sigma_x^2}{\Delta^2}}\right], \tag{2.56}$$

[10] Though x_k^q is not an integer, it lies on scaled integer lattice, and is sometimes nevertheless called *integer* in the literature. This slightly abusive term will be used in the second part of this thesis.

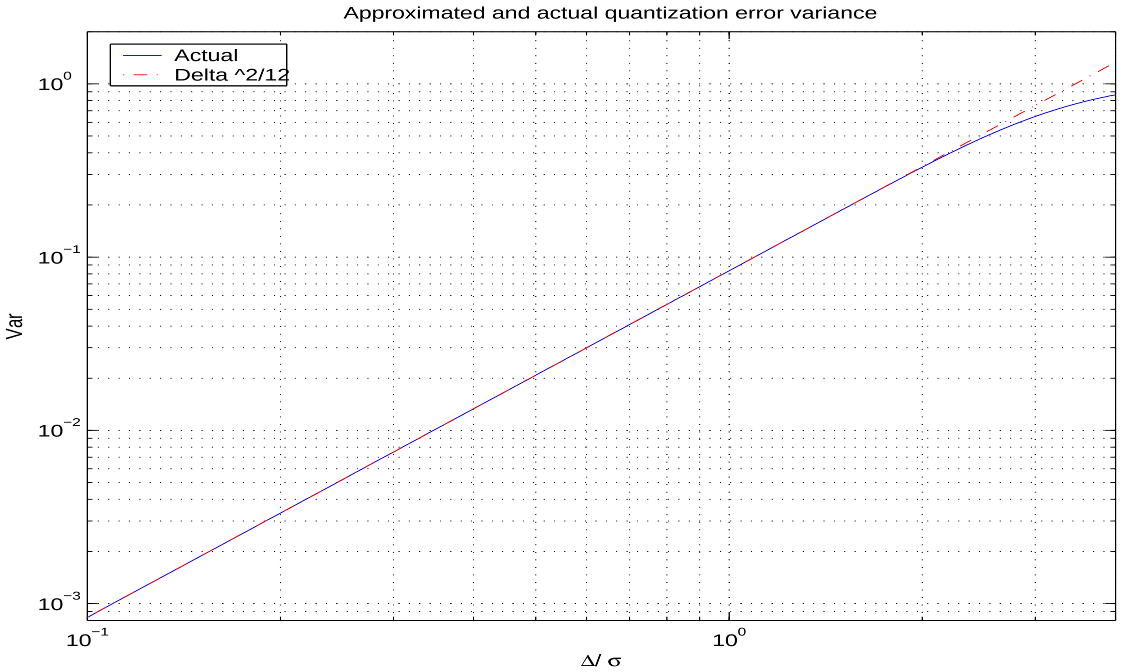


Figure 2.19: Comparison between actual distortion and $\frac{\Delta^2}{12}$ approximation for unbounded uniform quantization of a Gaussian r.v., $\sigma_x^2 = 1$.

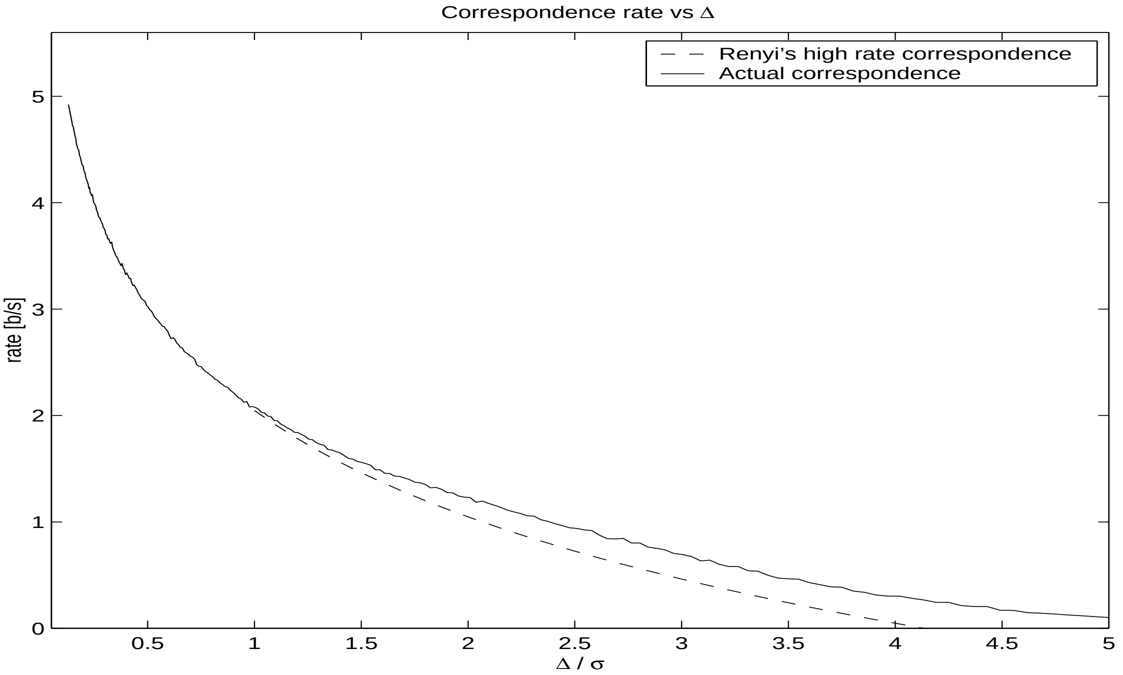


Figure 2.20: Estimated, and Rényi's correspondence Δ/σ *vs* actual rate for a Gaussian r.v..

and the correlation between the input and the quantization error by

$$\mathrm{E}\left(\tilde{x}x\right) = -2\sigma_x^2 \sum_{k=1}^{\infty} (-1)^k e^{-\frac{2\pi k^2 \sigma_x^2}{\Delta^2}} \tag{2.57}$$

Figure 2.21 compares the actual variance $\mathrm{E}\left(x^q\right)^2$ as given by (2.56) to the $\frac{\Delta^2}{12}$ approximation. Again, the approximation fits well the reality up to $\Delta \approx 1.5$, which from fig. 2.19 corresponds to approximately 1.5 b/s.

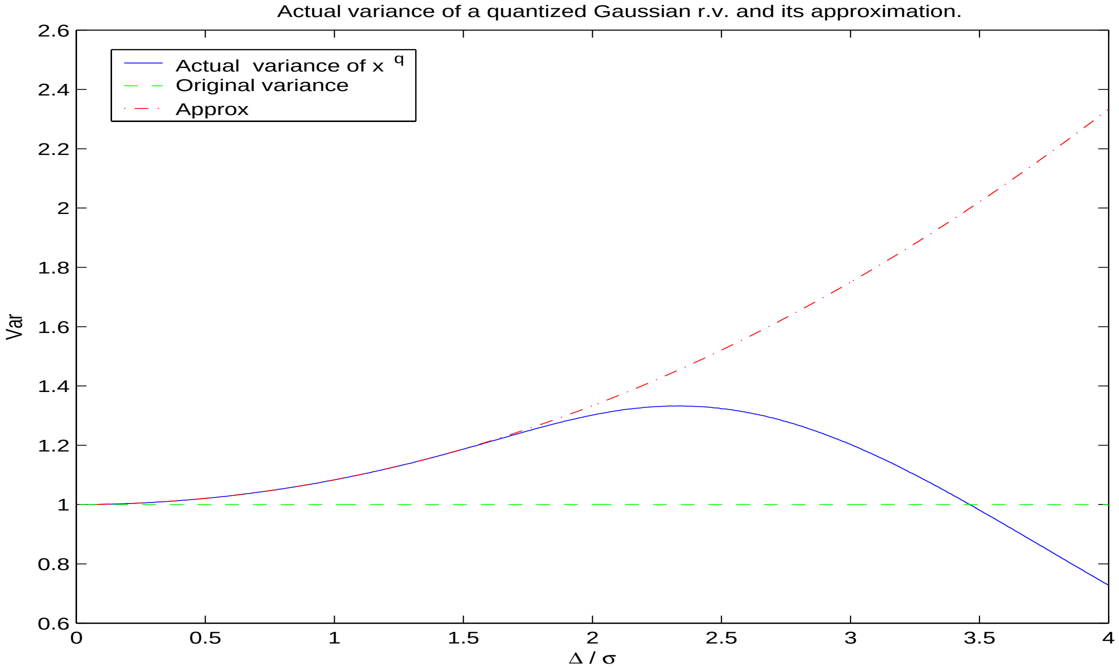


Figure 2.21: Comparison between actual and theoretic $\sigma_{x^q}^2$ for unbounded uniform quantization of a Gaussian r.v., $\sigma_x^2 = 1$.

In the case were jointly Gaussian zero mean r.v.s x_1 and x_2 with covariance $R_{\underline{xx}}$ are quantized with the same stepsize Δ, the quantization errors $\tilde{x}_1$ and $\tilde{x}_2$ verify

$$E(\tilde{x}_1\tilde{x}_2) = \frac{\Delta^2}{\pi^2} \sum_{l=1}^{\infty}\sum_{k=1}^{\infty} \frac{(-1)^{l+k}}{lk} e^{-\frac{2\pi^2}{\Delta^2}(l^2(R_{\underline{xx}})_{11}+k^2(R_{\underline{xx}})_{22})} \sinh\left(\frac{4\pi^2 lk(R_{\underline{xx}})_{12}}{\Delta^2}\right). \tag{2.58}$$

From the expression (22) of [72], applied to a roundoff quantizer, the correlation between quantized variables is given by

$$\mathrm{E}\left(x_1^q x_2^q\right) = (R_{\underline{xx}})_{12}(1+\delta) + \mu, \tag{2.59}$$

where

$$\delta = 2\left[\sum_{n_1>0} (-1)^{n_1} e^{-\frac{2n_1^2\pi^2(R_{\underline{xx}})_{11}}{\Delta^2}} + \sum_{n_2>0} (-1)^{n_2} e^{-\frac{2n_2^2\pi^2(R_{\underline{xx}})_{22}}{\Delta^2}}\right], \tag{2.60}$$

and

$$\mu = \sum_{n_1,n_2>0} (-1)^{n_1+n_2} \frac{\Delta^2}{n_1 n_2 \pi^2} e^{-\frac{2\pi^2(n_1(R_{\underline{x}\underline{x}})_{11}+n_2(R_{\underline{x}\underline{x}})_{22})}{\Delta^2}} \sinh\left(\frac{4\pi^2(R_{\underline{x}\underline{x}})_{12} n_1 n_2}{\Delta^2}\right). \qquad (2.61)$$

Similar conclusions regarding the accuracy of the classical approximation of respectively (2.57), (2.59) and (2.60), by respectively 0,1, and 0 can be drawn. Hence, for a Gaussian vector source $\underline{x}$, uniformly quantized with stepsize Δ we have from (2.55) and (2.58)

$$\mathrm{E}\,\tilde{x}\tilde{x}^T = \frac{\Delta^2}{12} I + A, \quad A \to 0 \text{ elementwise as } \Delta \to 0. \qquad (2.62)$$

From (2.56), (2.60) and (2.61), we have

$$\mathrm{E}\,\underline{x}^q \underline{x}^q T = R_{\underline{x}\underline{x}} + \frac{\Delta^2}{12} I + B, \quad B \to 0 \text{ elementwise as } \Delta \to 0. \qquad (2.63)$$

Moreover, the above numerical results and those of [70, 72] suggest that the elements of A and B may be negligible for Δ as large the standard deviations of the sources; this corresponds for entropy coded uniform quantizers to rates as low as approximately 1.5 b/s.

2.B Existence and unicity of the LDU factorization (2.8)

Lemma 1[73]: *Let R be an $N \times N$ nonsingular matrix, whose all principle submatrices are nonsingular. Then R can be written as*

$$R = L\ D\ U\ , \qquad (2.64)$$

where L (resp. U) is a lower (resp. upper) triangular matrix with diagonal entries equal to 1, and D is a diagonal matrix. Moreover, L, D and U are unique.

Since the covariance matrix $R_{\underline{x}\underline{x}}$ is positive definite, all its principle submatrices are positive definite also, and its LDU factorization (2.64) exists. Since now $R_{\underline{x}\underline{x}}$ is symmetric, transposing (2.64) yields

$$R_{\underline{x}\underline{x}} = U^T\ D\ L^T\ , \qquad (2.65)$$

where U^T (resp. L^T) is lower (resp. upper) triangular. Equation (2.65) represents then an LDU factorization of $R_{\underline{x}\underline{x}}$ which, from Lemma 1, is unique. Hence, $U = L^T$, which establishes the form of (2.8).

2.C Derivation of (2.40)

Let V denote a KLT of $R_{\underline{x}\underline{x}}$ and V' a KLT of $R_{\underline{x}\underline{x}} + {\sigma'}_q^2 I$.Then $V'(R_{\underline{x}\underline{x}} + {\sigma'}_q^2 I)V'^T = \Lambda'$, a diagonal matrix with i-th entry $\lambda'_i = \lambda_i + {\sigma'}_q^2$. Similarly, denote by L and L' the lower matrices involved in the LDU factorization of $R_{\underline{x}\underline{x}}$ and $R_{\underline{x}\underline{x}} + {\sigma'}_q^2 I$. We have as in (2.8) $LR_{\underline{x}\underline{x}}L^T = R_{\underline{y}\underline{y}}$, and, as in (2.33), $L'(R_{\underline{x}\underline{x}} + {\sigma'}_q^2 I)L'^T = R'_{\underline{y}\underline{y}} + {\sigma'}_q^2 I$, where $(R'_{\underline{y}\underline{y}})_{ii} = {\sigma'}_{y_i}^2$. As in (2.36), the variances ${\sigma'}_{y_i}^2$ may be written as $\sigma^2_{y_i} + \Delta_{\sigma^2_{y_i}}$, where $\Delta_{\sigma^2_{y_i}} = {\sigma'}^2_q||\overline{L_i}||^2$. Now, since both the causal and unitary transforms V' and L' are unimodular, we have

$$\begin{aligned} \det(R'_{\underline{yy}} + {\sigma'}^2_q I) &= \det\Lambda' \\ \prod_{i=1}^{N}\sigma^2_{y_i}\left(1 + \frac{{\sigma'}^2_q + \Delta_{\sigma^2_{y_i}}}{\sigma^2_{y_i}}\right) &= \left(\prod_{i=1}^{N}\lambda_i\right)\left(1 + \tfrac{{\sigma'}^2_q}{\lambda_i}\right) \\ \left(\prod_{i=1}^{N}\sigma^2_{y_i}\right)\left[1 + \sum_{i=1}^{N}\frac{{\sigma'}^2_q + \Delta_{\sigma^2_{y_i}}}{\sigma^2_{y_i}}\right] &\approx \left(\prod_{i=1}^{N}\lambda_i\right)\left[1 + \sum_{i=1}^{N}\frac{{\sigma'}^2_q}{\lambda_i}\right]. \end{aligned} \tag{2.66}$$

Since L and V are also unimodular, $\prod_{i=1}^{N}\lambda_i = \prod_{i=1}^{N}\sigma^2_{y_i}$, and we get

$$\begin{aligned} \sum_{i=1}^{N}\frac{{\sigma'}^2_q + \Delta_{\sigma^2_{y_i}}}{\sigma^2_{y_i}} &\approx \sum_{i=1}^{N}\frac{{\sigma'}^2_q}{\lambda_i} \\ \sum_{i=1}^{N}\frac{\Delta_{\sigma^2_{y_i}}}{\sigma^2_{y_i}} &\approx {\sigma'}^2_q\sum_{i=1}^{N}\left(\frac{1}{\lambda_i} - \frac{1}{\sigma^2_{y_i}}\right) \\ &\approx \sigma^2_q\sum_{i=1}^{N}\left(\frac{1}{\lambda_i} - \frac{1}{\sigma^2_{y_i}}\right) \end{aligned} \tag{2.67}$$

The required quantity for the distortion (2.38) and the coding gain (2.28) is the product

$$\begin{aligned} \prod_{i=1}^{N}{\sigma'}^2_{y_i} &= \prod_{i=1}^{N}\sigma^2_{y_i} + \Delta_{\sigma^2_{y_i}} \\ &\approx \left(\prod_{i=1}^{N}\sigma^2_{y_i}\right)\left[1 + \tfrac{\Delta_{\sigma^2_{y_i}}}{\sigma^2_{y_i}}\right] \\ &\approx \left(\prod_{i=1}^{N}\sigma^2_{y_i}\right)\left[1 + \sigma^2_q\sum_{i=1}^{N}\left(\frac{1}{\lambda_i} - \frac{1}{\sigma^2_{y_i}}\right)\right]. \end{aligned} \tag{2.68}$$

The distortion (2.38) becomes

$$\begin{aligned} \tfrac{1}{N}\,\mathrm{E}\,||\underline{\tilde{y}}||'^2 &= c2^{-2r}\left(\prod_{i=1}^{N}{\sigma'}^2_{y_i}\right)^{\frac{1}{N}} \\ &\approx c2^{-2r}(\det R_{\underline{xx}})^{\frac{1}{N}}\left[1 + \tfrac{\sigma^2_q}{N}\sum_{i=1}^{N}\left(\frac{1}{\lambda_i} - \frac{1}{\sigma^2_{y_i}}\right)\right]. \end{aligned} \tag{2.69}$$

Now the coding gain (2.28) becomes

$$\begin{aligned} G_L^{(1)} &= \left(\frac{\prod\limits_{i=1}^{N} \sigma_{x_i}^2}{\prod\limits_{i=1}^{N} {\sigma'}_{y_i}^2} \right)^{\frac{1}{N}} \\ &\approx \frac{\left(\prod\limits_{i=1}^{N} \sigma_{x_i}^2 \right)^{\frac{1}{N}}}{\left(\prod\limits_{i=1}^{N} \sigma_{y_i}^2 \right)^{\frac{1}{N}} \left[1+\sigma_q^2 \sum\limits_{i=1}^{N} \left(\frac{1}{\lambda_i} - \frac{1}{\sigma_{y_i}^2} \right) \right]^{\frac{1}{N}}} \\ &\approx \left(\frac{\prod\limits_{i=1}^{N} \sigma_{x_i}^2}{\prod\limits_{i=1}^{N} \sigma_{y_i}^2} \right)^{\frac{1}{N}} \left(1 - \frac{\sigma_q^2}{N} \sum_{i=1}^{N} \left[\frac{1}{\lambda_i} - \frac{1}{\sigma_{y_i}^2} \right] \right) , \end{aligned} \tag{2.70}$$

which with (2.10) yields (2.40).

Chapter 3

A High Resolution Analysis of Idealized Backward Adaptive Coding Schemes

In a backward adaptive transform coding framework, we compare the optimal unitary approach (Karhunen-Loève Transform, KLT) to the optimal causal approach (Lower-Diagonal-Upper, LDU). When the statistics of the source are known, the previous chapter showed that both coding schemes present the same coding gain at high rates. The purpose of this chapter is to analytically model the behavior of these two transformations when the ideal transform coding scheme gets perturbed, that is, when only a perturbed value $R_{\underline{xx}} + \Delta R$ *of* $R_{\underline{xx}}$ *is known at the encoder. This estimate is used to compute both the transforms and the bit assignment. This case is of interest in backward adaptive transform coding schemes: it avoids transmitting the updates of the signal-dependent transformations and bit assignment parameters as side information, and thereby avoids the corresponding excess bitrate. In backward adaptive structures,* ΔR *is due to two noise sources : estimation noise (finite set of available data at the encoder) and quantization noise (quantized data at the decoder). Furthermore, not only the transformation itself gets perturbed, but also the bit assignment. In this framework, theoretical expressions for the coding gains in both unitary and causal cases are derived, under several simplifying assumptions: high rate, Gaussianity of the signals to be quantized, same operational rate-distortion function of the scalar quantizers, optimal bit assignment, and additive uncorrelated white noise. Finally, simulations results are presented.*

3.1 Introduction

Backward adaptive coding schemes generally deal with non- or locally- stationary signals. In a transform coding framework, sending the updates of the signal-dependent transformation and bit assignment as side information may cause a considerable overhead for the overall bitrate. This problem is related to the general problem of *universal lossy quantization.* Universality corresponds here to the ability of a system which has no *a priori* knowledge of the source, to achieve the same rate-distortion performance as a system designed with that knowledge[1]. Very few works investigate the feasibility of universal transform codes in the literature. The work [54] is closely related to ours and will be further discussed in chapter 4. Besides, some techniques were proposed in [75, 76], which rely on so-called *two-stages* codes: the first stage codes the identity of the code that will be used to code the data; the second stage codes the data with the previously chosen code. In [76], a pair (KLT; bit allocation) is chosen among a codebook of transformations and bit allocations pairs; the index of the chosen pair is sent as side-information to the decoder. This type of method is universal in the sense that it allows one to code with the best transform and bit allocation any source among a particular class. The methods investigated in the present are different in the sense that they do not rely on "universal codebooks" of any kind. Instead of choosing among several precomputed transforms and bit allocations, we desire the encoder and decoder to compute these parameters using previously decoded data only. This approach is computationally more expensive, but does not require any side-information.

The backward adaptive transform coding scheme considered in the present work should therefore require that neither the transformation nor the parameters of the bit assignment are transmitted to the decoder. We assume consequently that the transform coding scheme is based on $\widehat{R}_{\underline{xx}} = R_{\underline{xx}} + \Delta R$ instead of $R_{\underline{xx}}$. $R_{\underline{xx}}$ is the unknown covariance matrix of a (possibly locally) stationary Gaussian vectorial process $\underline{x}$, and $\widehat{R}_{\underline{xx}}$ is the corresponding estimate available at both the encoder and the decoder. In this case, the computed transformation will be $\widehat{T} = T + \Delta T$, and the distortion will be proportional to the variances of the signals transformed by means of $\widehat{T}$ instead of T, say ${\sigma'}^2_{y_i}$. Moreover, the bits r_i should be attributed on the basis of estimates of the variances available at both encoder and decoder also. These estimates are $(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T})_{ii}$, where $(.)_{ii}$ denotes the ith diagonal element of $(.)$, which yields

$$\widehat{r}_i = r + \frac{1}{2}\log_2 \frac{(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T)_{ii}}{(\prod_{i=1}^N(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T)_{ii})^{\frac{1}{N}}}. \tag{3.1}$$

We obtain therefore the following measure of distortion for a transformation $\widehat{T}$ based on $\widehat{R}_{\underline{xx}}$:

$$\mathrm{E}\,||\tilde{\underline{y}}||^2_{\widehat{T}} = \mathrm{E}\sum_{i=1}^N c2^{-2\widehat{r}_i}{\sigma'}^2_{y_i} = \mathrm{E}\sum_{i=1}^N c2^{-2[r + \frac{1}{2}\log_2 \frac{(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T)_{ii}}{(\prod_{i=1}^N(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T)_{ii})^{\frac{1}{N}}}]}{\sigma'}^2_{y_i}. \tag{3.2}$$

[1] Different kind of universality for lossy coding, or with a fidelity criterion are defined in [74].

where the expectation is w.r.t. ΔR in case it is non-deterministic[2].

Several assumptions are implicitly made by the above description.

Firstly, we assume a Gaussian source model.

Secondly, the rate must be sufficiently high. The bit assignment mechanism (3.1) neglects the fact that $\widehat{r}_i$ can be noninteger and negative, which would happen for low values of the average bitrate budget r; or, even at higher values of r, for too low values of some transform coefficients y_i.

Thirdly, the expression (3.2) assumes that the quantizers' operational distortion-rate laws are of the form $c2^{-2r_i}\sigma'^2_{y_i}$, which assumes, besides high rates (independence of c w.r.t. to r_i) and significance of all transform coefficient (they are assigned nonzero r_i), that all the transform signals have the same p.d.f.s. For jointly Gaussian scalar sources x_i composing the vectorial source $\underline{x}$, this assumption is clearly true for the transform signals obtained by means of a KLT. In the case of a causal transform however, this is not rigorous even at high rates, because the prediction residuals y_i contain a quantized component through the closed loop prediction (Sec. 2.5.2). We shall therefore assume that this perturbation is small enough at high rates for the shape of the p.d.f. of all y_i to be accurately approximated by a Gaussian p.d.f..

Fourthly, we assume that the effects of quantization is to introduce a uncorrelated white noise with variance $c2^{-2\widehat{r}_i}\sigma'^2_{y_i}$.

Finally, in the case where estimation noise is involved, the vectors to be coded are assumed to be i.i.d. This restricts the scope of our analysis, but may be the case if the sampling period of the scalar signals is high in comparison with their typical correlation time.

The expected distortion (3.2) is thus a model subject to these assumptions, which make however analytical derivations possible. The goal of this work is then to provide, and compare in this framework the distortions and the corresponding coding gains for the KLT and the LDU, in three cases. In a first case (section 3.2), ΔR is caused by a quantization noise: the coding scheme is based on the statistics of the quantized data. In a second case, ΔR corresponds in section 3.3 to an estimation noise : the coding scheme is based on an estimate of $R_{\underline{x}\underline{x}}$ due to a finite amount of K vectors : $\widehat{R}_{\underline{x}\underline{x}} = \frac{1}{K}\sum_{i=1}^{K}\underline{x}_i\underline{x}_i^T$. Finally, both influences of quantization and estimation noise are analyzed in section 3.4. Numerical simulations are presented in section 3.5; the last section discusses the results and draws some conclusions.

3.2 Quantization Effects on the Coding Gains

Suppose we compute the transformation on the basis of quantized data. The statistics of the quantized data is therefore assumed to be perfectly known in this section. In other words, we assume that an infinite number of quantized vectors $\underline{x}_i^q$ is available at the decoder so that $R_{\underline{x}^q\underline{x}^q}$ is known. Under the above assumptions,

[2] The sign $=$ will be used, as in (3.2), along in the derivations though this equality is only correct asymptotically (w.r.t. to the rate) and under the discussed assumptions; the sign $\approx$ will be used when the original expression (3.2) will be replaced by its approximation based on the dominant perturbation effects.

$\Delta R = E\underline{\tilde{x}}\underline{\tilde{x}}^T = \sigma_q^2 I$. Thus, the distortion (3.2) becomes

$$\mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{T},q} = \sum_{i=1}^{N} c 2^{-2[r + \frac{1}{2}\log_2 \frac{(\widehat{T} R_{\underline{x}^q \underline{x}^q} \widehat{T}^T)_{ii}}{(\prod_{i=1}^{N} (\widehat{T} R_{\underline{x}^q \underline{x}^q} \widehat{T}^T)_{ii})^{\frac{1}{N}}}]} {\sigma'}^2_{y_i}, \tag{3.3}$$

where q refers to quantization. Expression (3.3) may now be evaluated for $\widehat{T} = I, \widehat{V}$ and $\widehat{L}$.

3.2.1 Identity Transformation

In this case, the number of bits attributed to the quantizer Q_i is

$$\widehat{r}_i = r + \frac{1}{2}\log_2 \frac{(R_{\underline{x}^q \underline{x}^q})_{ii}}{(\prod_{i=1}^{N} (R_{\underline{x}^q \underline{x}^q})_{ii})^{\frac{1}{N}}}, \tag{3.4}$$

and the variance ${\sigma'}^2_{y_i}$ are indeed $(R_{\underline{x}\underline{x}})_{ii}$. The distortion (3.3) becomes

$$\begin{aligned} \mathrm{E}\,||\underline{\tilde{y}}||^2_{I,q} &= \sum_{i=1}^{N} c 2^{-2[r + \frac{1}{2}\log_2 \frac{(R_{\underline{x}^q \underline{x}^q})_{ii}}{(\prod_{i=1}^{N} (R_{\underline{x}^q \underline{x}^q})_{ii})^{\frac{1}{N}}}]} (R_{\underline{x}\underline{x}})_{ii} \\ &= \sum_{i=1}^{N} c 2^{-2r} \left(\det \text{ diag}\{R_{\underline{x}^q \underline{x}^q}\}\right)^{\frac{1}{N}} \frac{(R_{\underline{x}\underline{x}})_{ii}}{(R_{\underline{x}^q \underline{x}^q})_{ii}}. \end{aligned} \tag{3.5}$$

The second equality comes from the fact that optimal bit assignment produces equal distortion on each component : suppose we compute the optimal bit assignment for an hypothetic signal with covariance matrix $R_{\underline{x}^q \underline{x}^q}$. Then we can write

$$\sum_{i=1}^{N} c 2^{-2[r + \frac{1}{2}\log_2 \frac{(R_{\underline{x}^q \underline{x}^q})_{ii}}{(\prod_{i=1}^{N} (R_{\underline{x}^q \underline{x}^q})_{ii})^{\frac{1}{N}}}]} (R_{\underline{x}^q \underline{x}^q})_{ii} = N c 2^{-2r} \left(\det \text{ diag}\{R_{\underline{x}^q \underline{x}^q}\}\right)^{\frac{1}{N}} = N\sigma^2_{q'}, \tag{3.6}$$

where all the terms $c2^{-2[r+\frac{1}{2}log_2 \frac{(R_{\underline{x}^q \underline{x}^q})_{ii}}{(\prod_{i=1}^{N} (R_{\underline{x}^q \underline{x}^q})_{ii})^{\frac{1}{N}}}]} (R_{\underline{x}^q \underline{x}^q})_{ii}$ are equal to some $\sigma^2_{q'}$ (each term equals the arithmetic mean of the right hand side term in (3.6)). Replacing $c2^{-2\widehat{r}_i}$ by $\sigma^2_{q'}/(R_{\underline{x}^q \underline{x}^q})_{ii}$ in the first equality of (3.5) gives the second equality.

Now, by writing $R_{\underline{x}^q \underline{x}^q} = R_{\underline{x}\underline{x}} + \sigma_q^2 I$ in (3.5), one can check that

$$\sum_{i=1}^{N} \frac{(R_{\underline{x}\underline{x}})_{ii}}{(R_{\underline{x}^q \underline{x}^q})_{ii}} = \text{tr}\left\{\left(I + \sigma_q^2(\text{ diag}\, R_{\underline{x}\underline{x}})^{-1}\right)^{-1}\right\}, \tag{3.7}$$

where tr denotes the trace operator, and also

$$\det\left(\text{diag}\, R_{\underline{x}^q \underline{x}^q}\right) = \det\left(\text{diag}\, R_{\underline{x}\underline{x}}\right) \det(I + \sigma_q^2(\text{ diag}\, R_{\underline{x}\underline{x}})^{-1}). \tag{3.8}$$

We obtain

$$\mathrm{E}\,||\underline{\tilde{y}}||^2_{I,q} = \mathrm{E}\,||\underline{\tilde{y}}||^2_{I} \frac{1}{N} (\det(I + \sigma_q^2(\text{ diag}\, R_{\underline{x}\underline{x}})^{-1}))^{\frac{1}{N}} \text{ tr}\left\{\left(I + \sigma_q^2(\text{ diag}\, R_{\underline{x}\underline{x}})^{-1}\right)^{-1}\right\}. \tag{3.9}$$

The distortion is slightly increased because the bits allocated on the basis of variances of quantized signals are not the optimal ones. An approximation of (3.9) up to the second order of the perturbation gives

$$\begin{aligned} \mathrm{E}\,\|\tilde{\underline{y}}\|^2_{I,q} &= c2^{-2r}(\det\ \mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{1/N}(\Pi_{i=1}^N(\tfrac{\sigma_q^2}{(R_{\underline{x}\underline{x}})_{ii}}))^{1/N}\sum_{i=1}^{N}(1+\frac{1}{(R_{\underline{x}\underline{x}})_{ii}})^{-1} \\ &\approx \mathrm{E}\,\|\tilde{\underline{y}}\|^2_I\left[1+\tfrac{\sigma_q^4}{N^2}(\tfrac{N-1}{2}\sum_{i=1}^{N}\frac{1}{(R_{\underline{x}\underline{x}})^2_{ii}}-\sum_{i=1}^{N}\sum_{j>i}\frac{1}{(R_{\underline{x}\underline{x}})_{ii}(R_{\underline{x}\underline{x}})_{jj}})\right]. \end{aligned} \tag{3.10}$$

3.2.2 KLT

As observed in [54] also, if V denotes a KLT of $R_{\underline{x}\underline{x}}$, then $V(R_{\underline{x}\underline{x}}+\sigma_q^2 I)V^T = \Lambda+\sigma_q^2 I = \Lambda^q$, and V is also a KLT of $R_{\underline{x}\underline{x}}+\sigma_q^2 I$. Thus, the perturbation term $\sigma_q^2 I$ on $R_{\underline{x}\underline{x}}$ does not change the backward adapted transformation: $\widehat{V} = V$. The variances of the transformed signals remain unchanged: ${\sigma'}^2_{y_i} = (VR_{\underline{x}\underline{x}}V^T)_{ii} = \lambda_i$. However, the decoder can only estimate the variances $(VR_{\underline{x}^q\underline{x}^q}V^T)_{ii} = \lambda_i+\sigma_q^2$, on the basis of which are assigned the bits $\widehat{r}_i$,

$$\widehat{r}_i = r+\frac{1}{2}\log_2\frac{(VR_{\underline{x}^q\underline{x}^q}V^T)_{ii}}{(\prod_{i=1}^N(VR_{\underline{x}^q\underline{x}^q}V)_{ii})^{\frac{1}{N}}}. \tag{3.11}$$

and the actual distortion becomes

$$\begin{aligned} \mathrm{E}\,\|\tilde{\underline{y}}\|^2_{V,q} &= \sum_{i=1}^{N} c2^{-2[r+\frac{1}{2}\log_2\frac{(VR_{\underline{x}^q\underline{x}^q}V^T)_{ii}}{(\prod_{i=1}^N(VR_{\underline{x}^q\underline{x}^q}V^T)_{ii})^{\frac{1}{N}}}]}(VR_{\underline{x}\underline{x}}V^T)_{ii} \\ &= \sum_{i=1}^{N} c2^{-2r}\left(\det\ \mathrm{diag}\,\{VR_{\underline{x}^q\underline{x}^q}V^T\}\right)^{\frac{1}{N}}\frac{(VR_{\underline{x}\underline{x}}V^T)_{ii}}{(VR_{\underline{x}^q\underline{x}^q}V^T)_{ii}}. \end{aligned} \tag{3.12}$$

Since $VR_{\underline{x}\underline{x}}V^T$ and $VR_{\underline{x}^q\underline{x}^q}V^T$ are diagonal, one shows that

$$\sum_{i=1}^{N}\frac{(VR_{\underline{x}\underline{x}}V^T)_{ii}}{(VR_{\underline{x}^q\underline{x}^q}V^T)_{ii}} = \mathrm{tr}\,\{\left(I+\sigma_q^2(R_{\underline{x}\underline{x}}^{-1})\right)^{-1}\} = \mathrm{tr}\,\{\left(I+\sigma_q^2(\Lambda)^{-1}\right)^{-1}\}. \tag{3.13}$$

Also,

$$\det\left(R_{\underline{x}^q\underline{x}^q}\right) = \det\left(R_{\underline{x}\underline{x}}\right)\det(I+\sigma_q^2(R_{\underline{x}\underline{x}}^{-1})). \tag{3.14}$$

Finally, the distortion for the KLT with quantization noise is

$$\mathrm{E}\,\|\tilde{\underline{y}}\|^2_{V,q} = \mathrm{E}\,\|\tilde{\underline{y}}\|^2_V\,\frac{1}{N}(\det(I+\sigma_q^2(\Lambda^{-1})))^{\frac{1}{N}}\,\mathrm{tr}\,\{\left(I+\sigma_q^2(\Lambda^{-1})\right)^{-1}\}. \tag{3.15}$$

Again, the increase in distortion comes from the perturbation occuring uppon the bit allocation mechanism. An expression approximating this distortion may be obtained by

$$\mathrm{E}\,\|\tilde{\underline{y}}\|^2_{V,q} = c2^{-2r}(\det\ \mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{\frac{1}{N}}\frac{1}{N}\left(\prod_{i=1}^{N}(1+\frac{\sigma_q^2}{\lambda_i})\right)^{\frac{1}{N}}\sum_{i=1}^{N}\left(1+\frac{\sigma_q^2}{\lambda_i}\right)^{-1}. \tag{3.16}$$

By developping the product and the sum in (3.16), it can be checked that the terms proportional to σ_q^2 vanish, so that

$$\left(\prod_{i=1}^{N}(1+\frac{\sigma_q^2}{\lambda_i})\right)^{\frac{1}{N}}\sum_{i=1}^{N}\left(1+\frac{\sigma_q^2}{\lambda_i}\right)^{-1} \approx N+\frac{N-1}{2N}\sum_i\frac{\sigma_q^4}{\lambda_i}-\frac{1}{N}\sum_{i=1}^{N}\sum_{j>i}\frac{\sigma_q^4}{\lambda_i\lambda_j}. \tag{3.17}$$

This leads for to the following distortion

$$\mathrm{E}\,||\underline{\tilde{y}}||_{V,q}^2 \quad\approx\quad \mathrm{E}\,||\underline{\tilde{y}}||_V^2\left[1+\tfrac{\sigma_q^4}{N^2}\left(\tfrac{N-1}{2}\sum_{i=1}^{N}\frac{1}{\lambda_i^2}-\sum_{i=1}^{N}\sum_{j>i}\frac{1}{\lambda_i\lambda_j}\right)\right] \tag{3.18}$$

Using (3.9) and (3.15), the corresponding expression for the coding gain in the unitary case with quantization noise is

$$G_{V,q}=G_{TC}^{(0)}\frac{(\det(I+\sigma_q^2(\operatorname{diag}R_{\underline{x}\underline{x}})^{-1}))^{\frac{1}{N}}\operatorname{tr}\{\left(I+\sigma_q^2(\operatorname{diag}R_{\underline{x}\underline{x}})^{-1}\right)^{-1}\}}{(\det(I+\sigma_q^2(\Lambda^{-1})))^{\frac{1}{N}}\operatorname{tr}\{\left(I+\sigma_q^2(\Lambda^{-1})\right)^{-1}\}}, \tag{3.19}$$

which, with (3.10) and (3.18), can be approximated as

$$G_{V,q}\approx G_{TC}^{(0)}\left[1+\frac{\sigma_q^4}{N^2}\left(\frac{N-1}{2}\sum_{i=1}^{N}(\frac{1}{(R_{\underline{x}\underline{x}})_{ii}^2}-\frac{1}{(\lambda_i)^2})-\sum_{i=1}^{N}\sum_{j>i}(\frac{1}{(R_{\underline{x}\underline{x}})_{ii}(R_{\underline{x}\underline{x}})_{jj}}-\frac{1}{\lambda_i\lambda_j})\right)\right]. \tag{3.20}$$

The perturbation effect w.r.t. to the ideal case is only due to the perturbation uppon the bit assignment, and appears to be weak (since it is proportional σ_q^4).

3.2.3 LDU

In the causal case, the coder computes a transformation $\widehat{L}=L'$ such that $L'R_{\underline{x}^q\underline{x}^q}L^{'T}=R'_{\underline{yy}}$. $R'_{\underline{yy}}$ is the diagonal matrix of the estimated variances involved in the bit allocation (L' and $R'_{\underline{yy}}$ are both available to the decoder). In this case, the difference vector $\underline{y}$ is $\underline{x}-\overline{L'}\underline{x}^q$. By the analysis of chapter 2, the quantization noise is filtered by the rows of $\overline{L'}$, see Figure 2.

Note that in this case $\mathrm{E}\,||\underline{\tilde{x}}||_{L',q}^2$ still equals $\mathrm{E}\,||\underline{\tilde{y}}||_{L',q}^2$, since $\underline{\tilde{x}}=\underline{x}^q-\underline{x}=\underline{y}^q+\overline{L'}\underline{x}^q-\underline{x}=\underline{\tilde{y}}$. Regarding the estimates of the rates, they are computed by

$$\widehat{r}_i=r+\frac{1}{2}\log_2\frac{(L'R_{\underline{x}^q\underline{x}^q}L^{'T})_{ii}}{(\prod_{i=1}^{N}(L'R_{\underline{x}^q\underline{x}^q}L'^T)_{ii})^{\frac{1}{N}}}. \tag{3.21}$$

It was shown in section 2.5 that the actual variances of the signals y_i obtained by means of L' and quantized x_I, q may be approximated as $(L'R_{\underline{x}^q\underline{x}^q}L^{'T}-\sigma_q^2I)_{ii}$ (see (2.34)). Using (3.3), the distortion $\mathrm{E}\,||\underline{\tilde{y}}||_{L',q}^2$ is then approximately given by

$$\begin{aligned}\mathrm{E}\,||\underline{\tilde{y}}||_{L',q}^2 &= c2^{-2[r+\frac{1}{2}\log_2\frac{(L'R_{\underline{x}^q\underline{x}^q}L^{'T})_{ii}}{(\prod_{i=1}^{N}(L'R_{\underline{x}^q\underline{x}^q}L'^T)_{ii})^{\frac{1}{N}}}]}\sum_{i=1}^{N}(L'R_{\underline{x}^q\underline{x}^q}L^{'T}-\sigma_q^2I)_{ii} \\ &\approx \sum_{i=1}^{N}c2^{-2r}\left(\det\ \operatorname{diag}\{L'R_{\underline{x}^q\underline{x}^q}L^{'T}\}\right)^{\frac{1}{N}}\left(1-\frac{(\sigma_q^2I)_{ii}}{(L'R_{\underline{x}^q\underline{x}^q}L'^T)_{ii}}\right).\end{aligned} \tag{3.22}$$

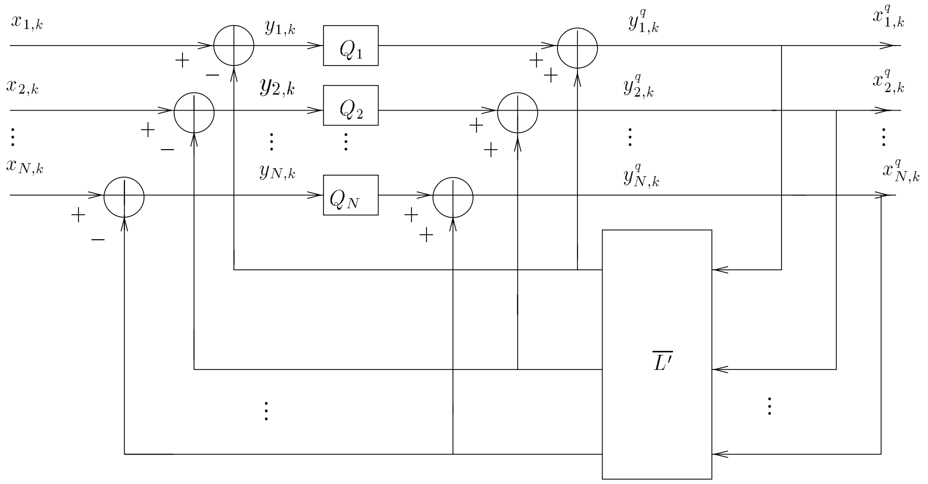


Figure 3.1: Backward adaptation of the causal transform.

Since the transformation L' is unimodular, the determinant in the previous expression equals the determinant in (3.14). The sum in (3.22) may be written as $\mathrm{tr}\{(I-\sigma_q^2(L'R_{\underline{x}^q\underline{x}^q}L'^T)^{-1})\} = \mathrm{tr}\{(I-\sigma_q^2 R'_{\underline{yy}}{}^{-1})\}$. Thus (3.22) becomes

$$\mathrm{E}\,||\underline{\tilde{y}}||^2_{L',q} = \mathrm{E}\,||\underline{\tilde{y}}||^2_L \frac{1}{N}(\det(I+\sigma_q^2(\Lambda^{-1})))^{\frac{1}{N}}\,\mathrm{tr}\left\{\left(I-\sigma_q^2(R'^{-1}_{\underline{yy}})\right)\right\}. \tag{3.23}$$

The increase in distortion comes not only from the perturbation occuring on the bit allocation mechanism but also from the filtering of the quantization noise. Up to the first order of perturbation, we obtain

$$\begin{aligned}\mathrm{E}\,||\underline{\tilde{y}}||^2_{(L',q} &= c2^{-2r}(\det\ \mathrm{diag}\,\{R_{\underline{xx}}\})^{\frac{1}{N}}\left(\prod_{i=1}^{N}(1+\frac{\sigma_q^2}{\lambda_i})\right)^{\frac{1}{N}}\sum_{i=1}^{N}\left(1-\sigma_q^2\frac{1}{(R'_{\underline{yy}})_{ii}}\right)\\ &\approx \mathrm{E}\,||\underline{\tilde{y}}||^2_V\left[1+\frac{\sigma_q^2}{N}\sum_{i=1}^{N}\left(\frac{1}{\lambda_i}-\frac{1}{\sigma_{y_i}^2}\right)\right],\end{aligned} \tag{3.24}$$

where the $\sigma_{y_i}^2$ correspond to optimal prediction error variances in absence of quantization noise. The corresponding expression for the coding gain is

$$G_{L',q} = G_{TC}^{(0)}\frac{(\det(I+\sigma_q^2(\,\mathrm{diag}\,\{R_{\underline{xx}}\})^{-1}))^{\frac{1}{N}}\,\mathrm{tr}\left\{\left(I+\sigma_q^2(\,\mathrm{diag}\,\{R_{\underline{xx}}\})^{-1}\right)^{-1}\right\}}{(\det(I+\sigma_q^2(\Lambda^{-1})))^{\frac{1}{N}}\,\mathrm{tr}\left\{\left(I-\sigma_q^2(R'^{-1}_{\underline{yy}})\right)\right\}}. \tag{3.25}$$

Up to the first order of perturbation we get,

$$G_{L',q} \approx G_{TC}^{(0)}\left[1-\frac{\sigma_q^2}{N}\sum_{i=1}^{N}\left(\frac{1}{\lambda_i}-\frac{1}{\sigma_{y_i}^2}\right)\right]. \tag{3.26}$$

The approximated expression (3.26) shows that the perturbation effects of the bit assignment mechanism are in the causal case negligible in comparison with those of the noise feedback; up to the first order of perturbation, this gain is similar to the gain obtained in (2.40), which corresponds to the case of noise feedback only, but with optimal bit allocation assumption. As in chapter 2, an interesting consequence of (3.26) is that the performance depends on the order in which the signals get decorrelated. It will be proved in chapter 5 that we should decorrelate the signals x_i in order of decreasing variance if we want $G_{L',q}$ to be maximized (see also Fig. 3.3 and 3.7 in the simulations section).

3.3 Estimation Noise

We analyze in this section the coding gains of a backward adaptive scheme based on an estimate of the covariance matrix $\widehat{R}_{\underline{xx}} = \frac{1}{K}\sum_{i=1}^{K} \underline{x}_i \underline{x}_i^T$. We assume independent identically distributed (i.i.d.) real vectors $\underline{x}_i$. Thus, the first and second order statistics of ΔR are known: one can show that ΔR is a zero mean Gaussian random variable, with (see section 3.B)

$$\mathrm{E}\,\mathrm{vec}(\Delta R)\,(\mathrm{vec}(\Delta R))^T \approx \frac{2}{K} R_{\underline{xx}} \otimes R_{\underline{xx}}, \tag{3.27}$$

where $\otimes$ denotes the Kronecker product. For each realization of ΔR, the coder computes a transformation $\widehat{T}$ which diagonalizes $\widehat{R}_{\underline{xx}}$: $\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T} = \widehat{R}_{\underline{yy}}$. The number of bits assigned to each component is therefore

$$\widehat{r}_i = r + \frac{1}{2}\log_2 \frac{(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T)_{ii}}{(\prod_{i=1}^{N}(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T)_{ii})^{\frac{1}{N}}}. \tag{3.28}$$

Now, the actual variances of the signals obtained by applying $\widehat{T}$ to $\underline{x}$ are $(\widehat{T}R_{\underline{xx}}\widehat{T}^T)_{ii}$. Note that in the causal case, $\underline{y} = I - \widehat{\overline{L}}\underline{x} = \widehat{L}\underline{x}$, so that $R'_{\underline{yy}} = \widehat{L}R_{\underline{xx}}\widehat{L}^T$. In the causal case, there is a qualitative difference with the previous section, where the quantization noise was filtered by the predictors of $\overline{L'}$. Here, the estimation noise does not perturb signals, but only transformations and bit assignments. The transformed signals are Gaussian for the three transformations, and the resulting distortion by estimating T and $R_{\underline{xx}}$ by means of K vectors is

$$\mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{T},K} = \mathrm{E}\sum_{i=1}^{N} c2^{-2[r+\frac{1}{2}\log_2 \frac{(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T)_{ii}}{(\prod_{i=1}^{N}(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T)_{ii})^{\frac{1}{N}}}]}(\widehat{T}R_{\underline{xx}}\widehat{T}^T)_{ii}. \tag{3.29}$$

3.3.1 Identity Transformation

With $\widehat{T} = I$, the resulting distortion is

$$\mathrm{E}\,||\underline{\tilde{y}}||^2_{I,K} = \mathrm{E}\sum_{i=1}^{N} c2^{-2[r+\frac{1}{2}\log_2 \frac{(\widehat{R}_{\underline{xx}})_{ii}}{(\prod_{i=1}^{N}(\widehat{R}_{\underline{xx}})_{ii})^{\frac{1}{N}}}]}(R_{\underline{xx}})_{ii}. \tag{3.30}$$

Using a similar analysis as in the previous section , we obtain

$$\begin{aligned} \mathrm{E}\,||\underline{\tilde{y}}||^2_{I,K} &= \mathrm{E}\,c2^{-2r}\left(\det\ \mathrm{diag}\,\{R_{\underline{xx}}\}\right)^{\frac{1}{N}}\left(\prod_{i=1}^{N}(1+\frac{(\Delta R)_{ii}}{(R_{\underline{xx}})_{ii}})\right)^{\frac{1}{N}}\sum_{i=1}^{N}\left(1+\frac{(\Delta R)_{ii}}{(R_{\underline{xx}})_{ii}}\right)^{-1} \\ &\approx \mathrm{E}\,||\underline{\tilde{y}}||^2_I\left(1+\ \mathrm{E}\ \tfrac{N-1}{2N^2}\sum_{i=1}^{N}(\frac{(\Delta R)_{ii}}{(R_{\underline{xx}})_{ii}})^2-\ \mathrm{E}\ \frac{1}{N^2}\sum_{i}\sum_{j>i}\frac{(\Delta R)_{ii}}{(R_{\underline{xx}})_{ii}}\frac{(\Delta R)_{jj}}{(R_{\underline{xx}})_{jj}}\right) \end{aligned} \tag{3.31}$$

With (3.27), the second expectation in (3.31) may be written as

$$\mathrm{E}\ \frac{N-1}{2N^2}\sum_{i=1}^{N}\left(\frac{(\Delta R)_{ii}}{(R_{\underline{xx}})_{ii}}\right)^2 \approx \frac{N-1}{2N^2}\sum_{i=1}^{N}\frac{2(R_{\underline{xx}})_{ii}^2}{K(R_{\underline{xx}})_{ii}^2} = \frac{N-1}{2N^2}\frac{2N}{K} = \frac{N-1}{NK}, \tag{3.32}$$

and the third expectation leads to

$$\begin{aligned}\mathrm{E}\ \tfrac{1}{N^2}\sum_i\sum_{j>i}\frac{(\Delta R)_{ii}}{(R_{\underline{xx}})_{ii}}\frac{(\Delta R)_{jj}}{(R_{\underline{xx}})_{jj}} &\approx \tfrac{2}{KN^2}\sum_i\sum_{j>i}\frac{(R_{\underline{xx}})_{ij}^2}{(R_{\underline{xx}})_{ii}(R_{\underline{xx}})_{jj}} \\ &\approx \tfrac{2}{KN^2}\|\triangleright\left((\mathrm{diag}\,\{R_{\underline{xx}}\})^{1/2}R_{\underline{xx}}(\mathrm{diag}\,\{R_{\underline{xx}}\})^{1/2}\right)\|^2\end{aligned} \tag{3.33}$$

where $\triangleright(A)$ denotes the stritcly lower triangular matrix made with the strictly lower triangular part of A. If D denotes $\mathrm{diag}\,\{R_{\underline{xx}}\}$, we obtain

$$\begin{aligned}\mathrm{E}\ \tfrac{1}{N^2}\sum_i\sum_{j>i}\frac{(\Delta R)_{ii}}{(R_{\underline{xx}})_{ii}}\frac{(\Delta R)_{jj}}{(R_{\underline{xx}})_{jj}} &\approx \tfrac{1}{KN^2}\left(\|D^{-\frac{1}{2}}R_{\underline{xx}}D^{-\frac{1}{2}}\|^2 - \underbrace{\|\,\mathrm{diag}\,\{D^{-\frac{1}{2}}R_{\underline{xx}}D^{-\frac{1}{2}}\|^2}_{N}\right) \\ &\approx \tfrac{1}{KN^2}(\,\mathrm{tr}\,\{R_{\underline{xx}}D^{-1}R_{\underline{xx}}D^{-1}\} - N).\end{aligned} \tag{3.34}$$

Finally, the expected distortion for Identity with estimation noise is, for sufficiently high K,

$$\mathrm{E}\,\|\underline{\tilde{y}}\|_{I,K}^2 \approx E\|\underline{\tilde{y}}\|_I^2\left(1+\frac{1}{K}[1-\frac{1}{N^2}\,\mathrm{tr}\,\{R_{\underline{xx}}(\,\mathrm{diag}\,\{R_{\underline{xx}}\})^{-1}R_{\underline{xx}}(\,\mathrm{diag}\,\{R_{\underline{xx}}\})^{-1}\}]\right). \tag{3.35}$$

3.3.2 KLT

In the unitary case, the expected distortion is

$$\mathrm{E}\,\|\underline{\tilde{y}}\|_{\widehat{V},K}^2 = \mathrm{E}\sum_{i=1}^{N} c2^{-2[r+\frac{1}{2}\log_2\frac{(\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T)_{ii}}{(\prod_{i=1}^{N}(\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T)_{ii})^{\frac{1}{N}}}]}(\widehat{V}R_{\underline{xx}}\widehat{V}^T)_{ii}. \tag{3.36}$$

Using an analysis similar to the previous subsection (see Sec. 3.C), the expected distortion for the KLT when the transformation is based on K vectors becomes, under high resolution assumption

$$\mathrm{E}\,\|\underline{\tilde{y}}\|_{\widehat{V},K}^2 \approx \mathrm{E}\,\|\underline{\tilde{y}}\|_{V}^2\left(1+\frac{N-1}{K}\left[\frac{1}{2}+\frac{1}{N}\right]\right). \tag{3.37}$$

The corresponding coding gain is

$$G_{\widehat{V},K} = \frac{\mathrm{E}\,\|\underline{\tilde{y}}\|_{I,K}^2}{\mathrm{E}\,\|\underline{\tilde{y}}\|_{\widehat{V},K}^2} \approx G_{TC}^{(0)}\left(1-\frac{1}{K}\left[\frac{\mathrm{tr}\,\{R(\,\mathrm{diag}\,\{R_{\underline{xx}}\})^{-1}R(\,\mathrm{diag}\,\{R_{\underline{xx}}\})^{-1}\}}{N^2}+\frac{N-1}{2}-\frac{1}{N}\right]\right). \tag{3.38}$$

3.3.3 LDU

As stated in the introduction of this section, the expected distortion with $\widehat{L}$ computed with $\widehat{R}_{\underline{xx}}$ is

$$\begin{aligned}\mathrm{E}\,\|\underline{\tilde{y}}\|_{\widehat{L},K}^2 &= \mathrm{E}\sum_{i=1}^{N} c2^{-2[r+\frac{1}{2}\log_2\frac{(\widehat{L}\widehat{R}_{\underline{xx}}\widehat{L}^T)_{ii}}{(\prod_{i=1}^{N}(\widehat{L}\widehat{R}_{\underline{xx}}\widehat{L}T)_{ii})^{\frac{1}{N}}}]}(\widehat{L}R_{\underline{xx}}\widehat{L}^T)_{ii} \\ &= \mathrm{E}\,c2^{-2r}\left(\det\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}\right)^{\frac{1}{N}}\sum_{i=1}^{N}\frac{(\widehat{L}R_{\underline{xx}}\widehat{L}^T)_{ii}}{(\widehat{L}\widehat{R}_{\underline{xx}}\widehat{L}^T)_{ii}},\end{aligned} \tag{3.39}$$

where we used a factorization similar to that used in (3.5). Now by the unimodularity property $\widehat{L}$, we can write the determinant in (3.39) as

$$\left(\det \widehat{V} \widehat{R}_{\underline{x}\underline{x}} \widehat{V}\right)^{\frac{1}{N}} = \det \widehat{R}_{\underline{x}\underline{x}} = \det(R_{\underline{x}\underline{x}}) \det(I + R_{\underline{x}\underline{x}}^{-1} \Delta R), \tag{3.40}$$

and since $\widehat{L}$ diagonalizes $\widehat{R}_{\underline{x}\underline{x}}$, we can write the sum in (3.39) as

$$\sum_{i=1}^{N} \frac{(\widehat{L} R_{\underline{x}\underline{x}} \widehat{L}^T)_{ii}}{(\widehat{L} \widehat{R}_{\underline{x}\underline{x}} \widehat{L}^T)_{ii}} = \text{tr}\{(I + R_{\underline{x}\underline{x}}^{-1} \Delta R)^{-1}\}. \tag{3.41}$$

Now because both causal LDU and unitary KLT are decorrelating and unimodular transforms, it can be checked that $\mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{L},K} = \mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{V},K}$ (comparing with the analysis in (3.65), the equality of the determinants comes from the unimodularity of the transformations $\widehat{L}$ and $\widehat{V}$, and the equality of the traces comes from their decorrelating property). The distortion and coding gain with estimation noise are then the same in the causal and the unitary cases; they may be approximated by (3.37) and (3.38) respectively.

3.4 Quantization and Estimation Noise

We arrive now to the most general case of this comparison between causal and unitary approaches. In presence of quantization and estimation noise, the bits should be attributed on the basis of $\widehat{R}_{\underline{x}^q\underline{x}^q} = \frac{1}{K}\sum_{i=1}^{K} \underline{x}_i^q \underline{x}_i^{qT}$. As in the previous section, we assume independent identically distributed real vectors $\underline{x}_i$. The estimated transform is $\widehat{T}$, such that $\widehat{T}\widehat{R}_{\underline{x}^q\underline{x}^q}\widehat{T}^T$ is a diagonal matrix, which corresponds to the estimated variances of the transformed signals. If we continue denoting by ${\sigma'}^2_{y_i}$ the actual variances of the transformed signals (obtained by applying $\widehat{T}$ to $\underline{x}_k$), the expected distortion (3.2), obtained with $\widehat{T}$ using K quantized vectors becomes

$$\mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{T},K,q} = \mathrm{E}\sum_{i=1}^{N} c2^{-2[r + \frac{1}{2}\log_2 \frac{(\widehat{T}\widehat{R}_{\underline{x}^q\underline{x}^q}\widehat{T}^T)_{ii}}{(\prod_{i=1}^{N}(\widehat{T}\widehat{R}_{\underline{x}^q\underline{x}^q}\widehat{T}^T)_{ii})^{\frac{1}{N}}}]} {\sigma'}^2_{y_i}, \tag{3.42}$$

where the subscripts ^{q} and ^{K} refer to the presence of quantization and estimation noise. Equation (3.42) must now be evaluated for Identity, KL and LDU transforms.

3.4.1 Identity Transformation

In this case the transformed signals y_i are indeed still Gaussian. With $\widehat{T} = I$ we obtain for (3.42), by writing $R_{\underline{x}\underline{x}} = R_{\underline{x}^q\underline{x}^q} - \sigma_q^2 I$,

$$\begin{aligned} \mathrm{E}\,||\underline{\tilde{y}}||^2_{I,K,q} = {} & \mathrm{E} \sum_{i=1}^{N} c2^{-2[r + \frac{1}{2}\log_2 \frac{(\widehat{R}_{\underline{x}^q\underline{x}^q})_{ii}}{(\prod_{i=1}^{N}(\widehat{R}_{\underline{x}^q\underline{x}^q})_{ii})^{\frac{1}{N}}}]} (R_{\underline{x}^q\underline{x}^q})_{ii} \\ & -\sigma_q^2\,\mathrm{E} \sum_{i=1}^{N} c2^{-2[r + \frac{1}{2}\log_2 \frac{(\widehat{R}_{\underline{x}^q\underline{x}^q})_{ii}}{(\prod_{i=1}^{N}(\widehat{R}_{\underline{x}^q\underline{x}^q})_{ii})^{\frac{1}{N}}}]} . \end{aligned} \tag{3.43}$$

The expected distortion for Identity transform with quantization and estimation noise may then (see sec. 3.D), for sufficiently high resolution and large K, be written as

$$\begin{aligned} \mathrm{E}\,||\underline{\tilde{y}}||^2_{I,K,q} \approx {} & \mathrm{E}\,||\underline{\tilde{y}}||^2_{I} \left(\det(I + \sigma_q^2(\,\mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{-1})\right)^{1/N} \\ & \times \left[1 + \tfrac{1}{K}\left[1 - \tfrac{1}{N^2}\,\mathrm{tr}\,\{R_{\underline{x}^q\underline{x}^q}(\,\mathrm{diag}\,R_{\underline{x}^q\underline{x}^q})^{-1}R_{\underline{x}^q\underline{x}^q}(\,\mathrm{diag}\,R_{\underline{x}^q\underline{x}^q})^{-1}\}\right] - \tfrac{\sigma_q^2}{N}\,\mathrm{tr}\,\{(\,\mathrm{diag}\,R_{\underline{x}^q\underline{x}^q})^{-1}\}\right] . \end{aligned} \tag{3.44}$$

3.4.2 KLT

The expected distortion (3.2) with quantization and estimation noise becomes, in the unitary case,

$$\mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{V},K,q} = \mathrm{E} \sum_{i=1}^{N} c2^{-2[r + \frac{1}{2}\log_2 \frac{(\widehat{V}\widehat{R}_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}{(\prod_{i=1}^{N}(\widehat{V}\widehat{R}_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii})^{\frac{1}{N}}}]} (\widehat{V}R_{\underline{x}\underline{x}}\widehat{V}^T)_{ii} . \tag{3.45}$$

After some computation (see sec. 3.E), we find for the expected distortion in the unitary case, when the transformation is based on K quantized vectors,

$$\mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{V},K,q} \approx \mathrm{E}\,||\underline{\tilde{y}}||^2_{K} \left(\det(I + \sigma_q^2(R_{\underline{x}\underline{x}})^{-1})\right)^{\frac{1}{N}} \left[1 + \frac{N-1}{K}\left[\frac{1}{2} + \frac{1}{N}\right] - \frac{\sigma_q^2}{N}\,\mathrm{tr}\,\{(R_{\underline{x}^q\underline{x}^q})^{-1}\}\right] , \tag{3.46}$$

for large K and under high resolution assumption. The corresponding expression for the coding gain is

$$\begin{aligned} G_{\widehat{V},K,q} = \frac{\mathrm{E}\,||\underline{\tilde{y}}||^2_{I,K,q}}{\mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{V},K,q}} \approx G_{TC}^{(0)} \frac{\left(\det(I + \sigma_q^2(\,\mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{-1})\right)^{1/N}}{\left(\det(I + \sigma_q^2(R_{\underline{x}\underline{x}})^{-1})\right)^{1/N}} \\ \times \frac{\left[1 + \frac{1}{K}(1 - \frac{1}{N^2}\,\mathrm{tr}\,\{R_{\underline{x}^q\underline{x}^q}(\,\mathrm{diag}\,R_{\underline{x}^q\underline{x}^q})^{-1}R_{\underline{x}^q\underline{x}^q}(\,\mathrm{diag}\,R_{\underline{x}^q\underline{x}^q})^{-1}\}) - \frac{\sigma_q^2}{N}\,\mathrm{tr}\,\{(\,\mathrm{diag}\,R_{\underline{x}^q\underline{x}^q})^{-1}\}\right]}{\left[1 + \frac{N-1}{K}(\frac{1}{2} + \frac{1}{N}) - \frac{\sigma_q^2}{N}\,\mathrm{tr}\,\{(R_{\underline{x}^q\underline{x}^q})^{-1}\}\right]} . \end{aligned} \tag{3.47}$$

3.4.3 LDU

In the causal case, an estimate $\widehat{L}'$ is computed, and the actual variances are ${\sigma'}^2_{y_i} = \mathrm{E}\,(\widehat{L}' R_{\underline{x}^q \underline{x}^q} \widehat{L}'^T - \sigma_q^2 I)_{ii}$. Thus, computing (3.42) when the transformation is based on K quantized vectors (for high K and under high resolution assumption) gives (see sec. 3.F)

$$\mathrm{E}\,\|\underline{\tilde{y}}\|^2_{\widehat{L}',K,q} = \mathrm{E} \sum_{i=1}^{N} c2^{-2[r + \frac{1}{2}\log_2 \frac{(\widehat{L}' \widehat{R}_{\underline{x}^q \underline{x}^q} \widehat{L}'^T)_{ii}}{(\prod_{i=1}^N (\widehat{L}' \widehat{R}_{\underline{x}^q \underline{x}^q} \widehat{L}'^T)_{ii})^{\frac{1}{N}}}]} (\widehat{L}' R_{\underline{x}^q \underline{x}^q} \widehat{L}'^T - \sigma_q^2 I)_{ii}, \tag{3.48}$$

which can be approximated as

$$\mathrm{E}\,\|\underline{\tilde{y}}\|^2_{\widehat{L}',K,q} = \mathrm{E}\,\|\underline{\tilde{y}}\|^2_{L} \left(\det(I + \sigma_q^2 (R_{\underline{x}\underline{x}})^{-1})\right)^{1/N} \left[1 + \frac{N-1}{K}\left[\frac{1}{2} + \frac{1}{N}\right] - \frac{\sigma_q^2}{N}\,\mathrm{tr}\,\{(R'_{\underline{y}\underline{y}})^{-1}\}\right], \tag{3.49}$$

The corresponding expression for the coding gain in the causal case can then be estimated as

$$\begin{aligned} G_{\widehat{L}',K,q} &= \frac{\mathrm{E}\,\|\underline{\tilde{x}}\|^2_{I,K,q}}{\mathrm{E}\,\|\underline{\tilde{y}}\|^2_{\widehat{L}',K,q}} \\ &= G_{TC}^{(0)} \frac{\left(\det(I + \sigma_q^2 (\,\mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{-1})\right)^{1/N}}{\left(\det(I + \sigma_q^2 (R_{\underline{x}\underline{x}})^{-1})\right)^{1/N}} \\ &\times \frac{\left[1 + \frac{1}{K}\left[1 - \frac{1}{N^2}\,\mathrm{tr}\,\{R_{\underline{x}^q \underline{x}^q} (\,\mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{q-1} R_{\underline{x}^q \underline{x}^q} (\,\mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{q-1}\}\right] - \frac{\sigma_q^2}{N}\,\mathrm{tr}\,\{(\,\mathrm{diag}\,R_{\underline{x}^q \underline{x}^q})^{-1}\}\right]}{\left[1 + \frac{N-1}{K}\left[\frac{1}{2} + \frac{1}{N}\right] - \frac{\sigma_q^2}{N}\,\mathrm{tr}\,\{(L' R_{\underline{x}^q \underline{x}^q} L'^T)^{-1}\}\right]}. \end{aligned} \tag{3.50}$$

It can be checked that the expressions (3.50) and (3.47) tend to (3.19) and (3.25) respectively as $K \to \infty$, and both to (3.38) as $\sigma_q^2 \to 0$.

3.5 Simulations

For the simulations, we generated real Gaussian i.i.d. vectors with covariance matrix $R_{\underline{x}\underline{x}_j} = H_j R_{AR1} H_j^T$, $j = 1, 2$. R_{AR1} denotes the covariance matrix of a first order autoregressive process with normalized cross correlation coefficient ρ. H_j is a diagonal matrix whose ith entry is $i^{1/3}$ for H_1 (increasing variances), and $(N - i + 1)^{1/3}$ (decreasing variances) for H_2. The goal of these numerical evaluations is to check if the distortion as described in (3.2) corresponds to the either exact, either approximated theoretical expressions which were derived in the three cases of quantization, estimation noise, and both. The following algorithms were therefore used check our analytical results.

3.5.1 Quantization Noise

For several rates (from 2 to 6 b/s), bit allocations and transforms ($\widehat{T} = I, L'$ and V respectively) were computed using $\widehat{R}_{\underline{x}\underline{x}} = R_{\underline{x}\underline{x}_j} + \sigma_q^2 I$, where $\sigma_q^2 = c2^{-2r} \det R_{\underline{x}\underline{x}_j}$ (that is, the distortion occuring in a high rate transform coding framework with optimal bit allocation). The choice of the constant is not relevant because (3.2) is very general; we chose $c = \frac{\pi e}{6}$. The bits to be allocated where then computed by (3.1), with $(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T)_{ii} = (R_{\underline{xx}})_{ii}$ for the Identity tranform, and by (3.11) and (3.21) in the unitary and causal cases respectively. The corresponding distortions where computed using (3.5), (3.12) and (3.22) respectively, with variances ${\sigma'}^2_{y_i} = (R_{\underline{xx}})_{ii}, (L'R_{\underline{x}^q\underline{x}^q}{L'}^T - \sigma_q^2 I)_{ii}$, and λ_i respectively. The resulting distortion where then computed to measure the coding gains which were compared with the theoretical expressions.

- In Figure 3.2, the coding gain with quantization noise is plotted for KLT (upper curves, full line) and LDU (lower curves, full line), for signals of decreasing variances, and with $\rho = 0.9$, $N = 4$. The theoretical exact expressions are given by (3.19) and (3.25), the corresponding curves are dotted. The theoretical approximated expressions are given by (3.20) and (3.26), and the corresponding curves are dashed.
- Figure 3.3 shows the influence of the variance ordering in the decorrelation process. The upper curve depicts the gain obtained with the causal approach by decorrelating the signals by decreasing order of variance (R_{xx_2}), and the lower curve by increasing order (R_{xx_1}).

It is checked that the expressions (3.19) and (3.25) are actually exact; the approximated expressions (3.20) and (3.26) match their exact counterparts above approximately 2.5 b/s.

3.5.2 Estimation Noise

In this case, estimates of the covariance matrix of the data was computed using K vectors by $\frac{1}{K}\sum_{i=1}^{K}\underline{x}_i\underline{x}_i^T$, $K = N, N+1, \cdots, 10^3$. For each estimate $\widehat{R}_{\underline{xx}}$, the transforms $\widehat{T} = \widehat{V}, \widehat{L}$ were computed so that $\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T$ is diagonal, and the bit allocations were computed using estimates of the variances $(\widehat{T}\widehat{R}_{\underline{xx}}\widehat{T}^T)_{ii}$. In order to evaluate the expected distortion (3.29), the sum in (3.29) was considered as a random variable, whose expectation was evaluated by Monte Carlo simulations. This was done for the Identity transform, in the causal and in the unitary case. The ratio of the corresponding distortions are the "Observed Coding Gain" in Figure 3.4. The corresponding theoretical expression is given by (3.38) (should be the same for the KLT and LDU because both transforms are decorrelating and unimodular). The coding gains in presence of estimation noise are compared in the figure for $N = 4$ and $\rho = 0.9$.

As expected, no difference can be noticed between the unitary and the causal case. Our calculations assumed small perturbations; it can be observed that the model matches the actual coding gain after a few tens of vectors.

3.5.3 Quantization and Estimation Noise

In this case, the quantized vectors were obtained for each rate r by perturbing the sets of i.i.d. Gaussian vector with uncorrelated white noise vectors with covariance matrix $\sigma_q^2 I = c2^{-2r}(\det R_{\underline{xx}})^{\frac{1}{N}} I$. For each set of available quantized vectors, an estimate of the covariance matrix of the data was computed by $\frac{1}{K}\sum_{i=1}^{K}\underline{x}_i^q\underline{x}_i^{q\,T}$, $K = N, N+1, \cdots, 10^3$. Again, for each estimate $\widehat{R}_{\underline{x}^q\underline{x}^q}$, the transforms $\widehat{T} = \widehat{V}, \widehat{L}$ were computed so that $\widehat{T}\widehat{R}_{\underline{x}^q\underline{x}^q}\widehat{T}^T$ is diagonal, and the bit allocations were computed using estimates of the variances $(\widehat{T}\widehat{R}_{\underline{x}^q\underline{x}^q}\widehat{T}^T)_{ii}$. In order to evaluate the expected distortion (3.42), the sum in (3.42) was considered as a random variable, whose expectation was evaluated by Monte Carlo simulations. This was done for the Identity transform, in the causal and in the unitary case. The ratio of the corresponding distortions are the "Observed Gains" of the following figures. The theoretical gains are given by (3.47) and (3.50).

- The coding gains in presence of estimation noise and quantization are compared for KLT and LDU (signals of decreasing variances) in figure 3.5 (res. 3.6), for $N = 8$ (resp. $N = 4$), $\rho = 0.9$ and a rate of 3 bits per sample. The observed behaviors of the transformation corresponds quite well to the theoretically predicted ones for $K \approx$ a few tens.

- The influence of the ordering of the signals for the same parameters as above is plotted in Figure 3.7.

In the limit if large K, the actual gains converge to the results obtained in the case where quantization noise only is considered (the estimation noise vanishes). The proposed models match the actual convergence behaviours in the causal and unitary cases after a few tens of decoded vectors. Finally, decorrelating the signals by order of decreasing variance appears the best strategy.

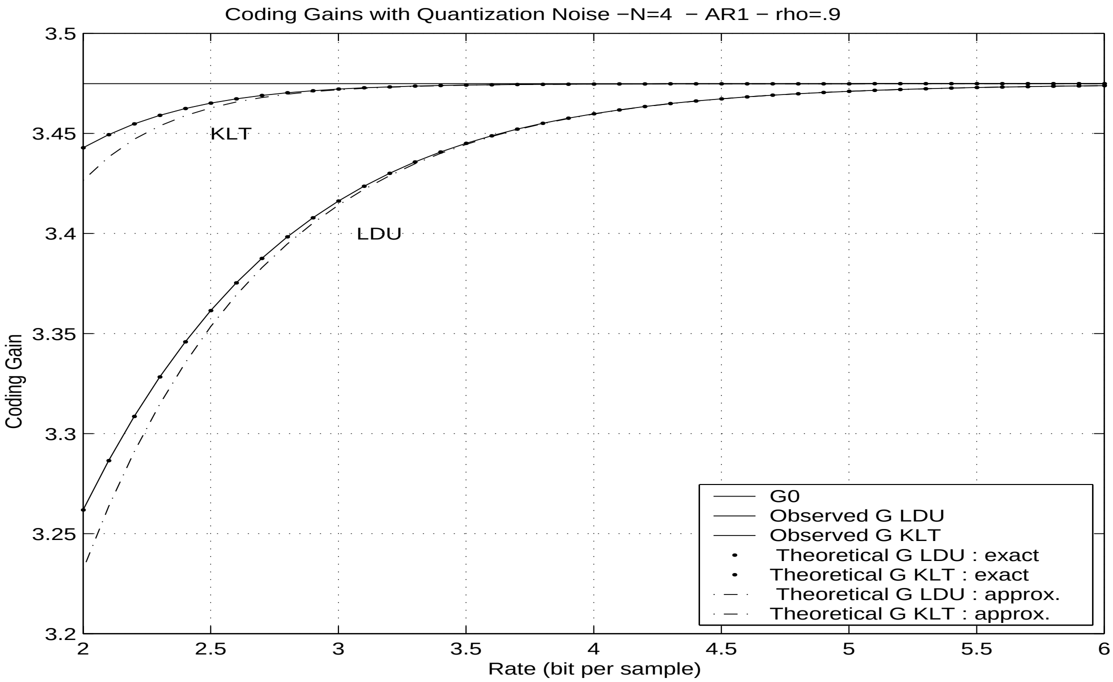


Figure 3.2: Coding Gains *vs* rate in bit/sample.

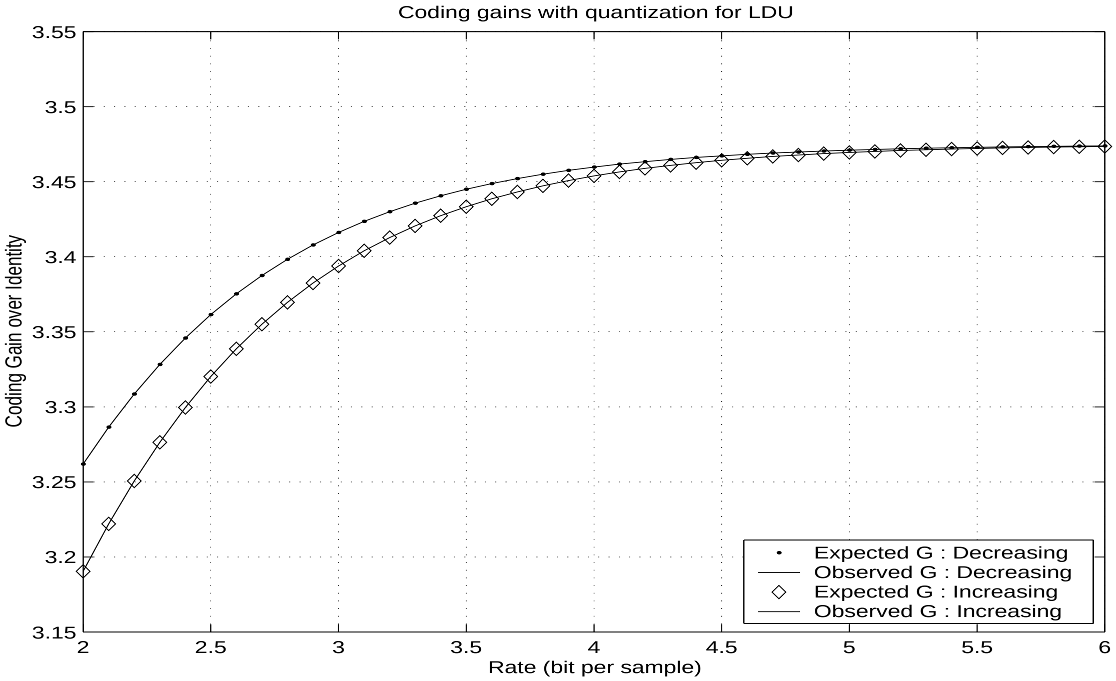


Figure 3.3: Influence of the ordering of the signals x_i.

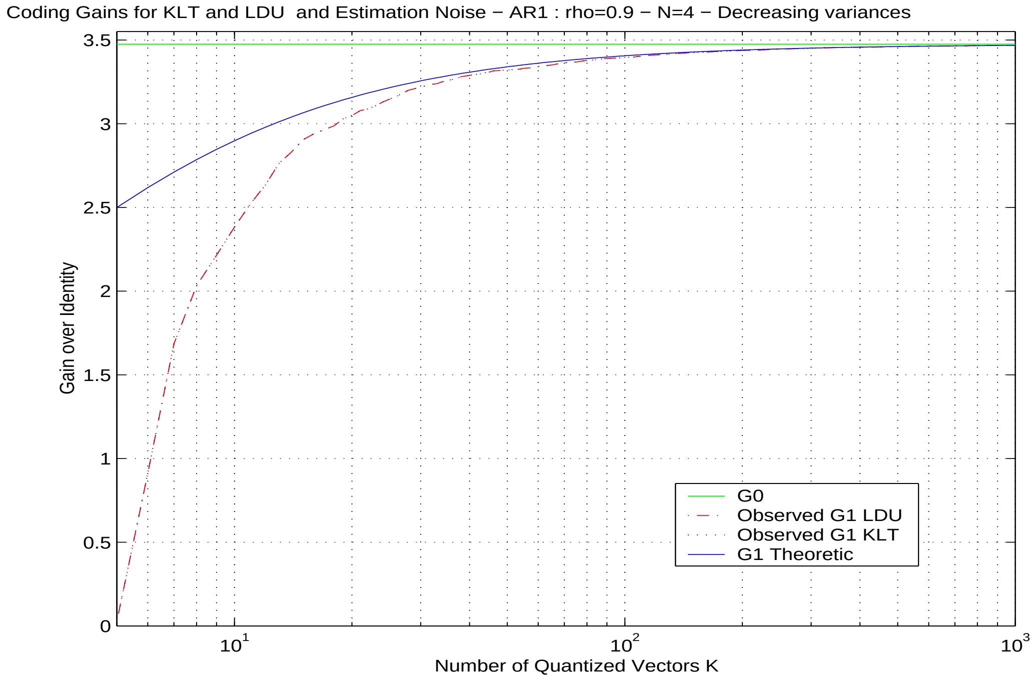


Figure 3.4: Gains for KLT and LDU with estimation noise.

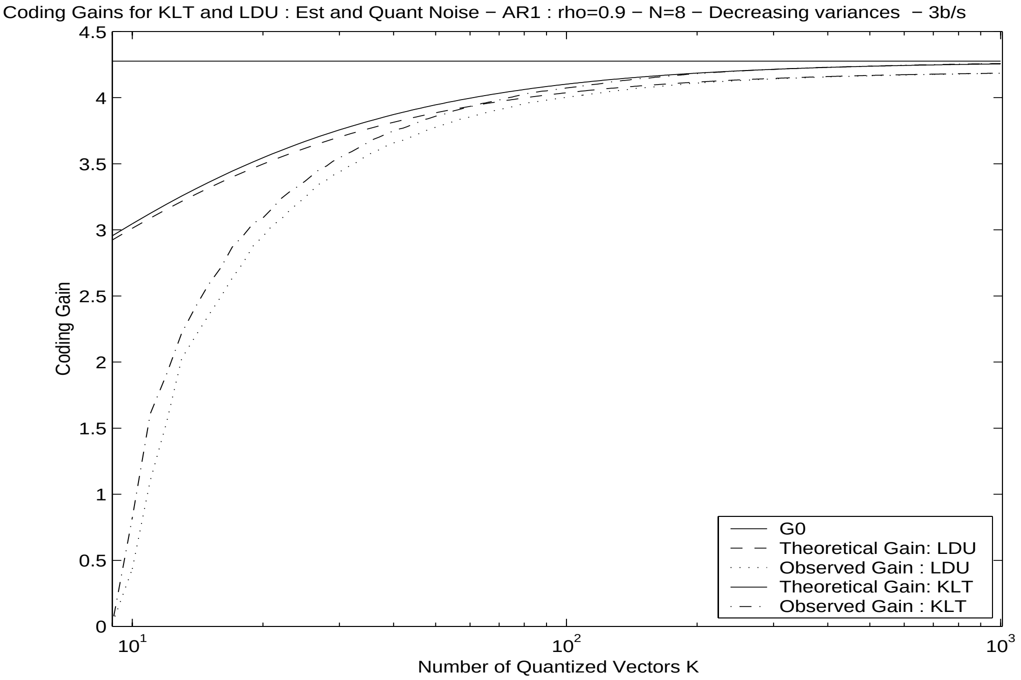


Figure 3.5: Gains for KLT and LDU $\rho = 0.9$, rate=3 b/s, $N = 8$.

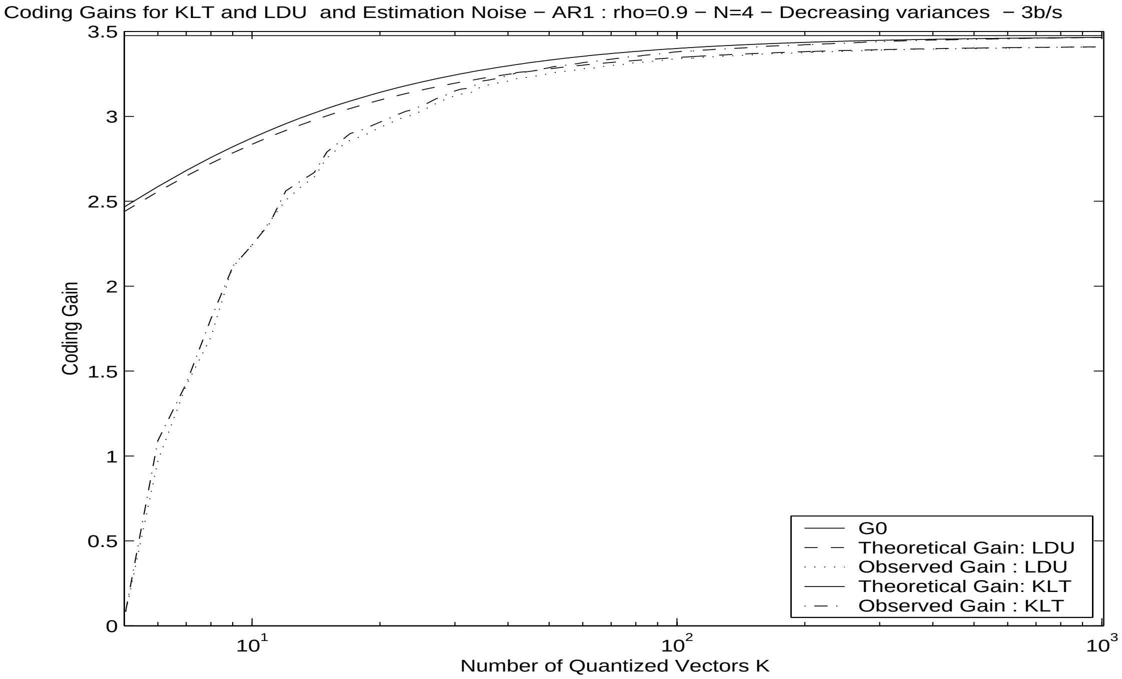


Figure 3.6: Gains for KLT and LDU $\rho = 0.9$. The rate is 3 b/s and $N = 4$.

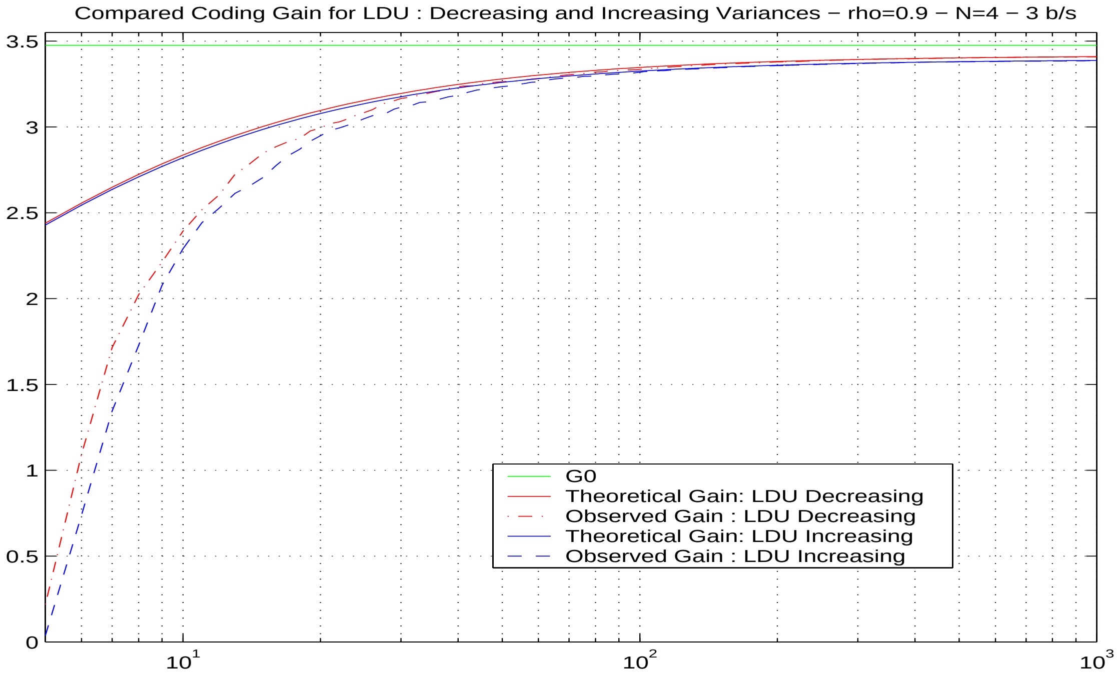


Figure 3.7: Ordering of the signals : Compared coding gains for LDU. The rate is 3 b/s and $N = 4$.

3.6 Discussion and Conclusions

This chapter has proposed an analytical model for the performance of causal and unitary backward adaptive transform coding schemes. The approach consisted in analyzing the effects of backward adaptation as perturbation effects; these effects perturb the ideal high rate transform coding coding framework by perturbing both the transforms' design and the bit assignment mechanism. The presented simulation results have shown that the resulting analytical description of the systems is fairly accurate. In particular, exact expression for the coding gain can be achieved as far as the quantization noise only is concerned. When estimation noise is accounted for, the proposed analysis reliably estimates the distortions and the corresponding coding gains after a few tens of decoded vectors.

The cost of the proposed analytical evaluation is the introduction of several simplifying assumptions. One may argue that some, if not all the considered assumptions may not be verified in practical cases. These objections are receivable: Gaussianity and independence of the source vectors may not correspond to real world sources. The additive quantization model is overly simplistic, if not incorrect, for quantizers which are not uniform. Finally, it is not clear how practical systems would actually realize the bit allocation procedure as assumed in (3.1).

These reflections led us to investigate further more particular, but practical backward adaptive systems. Some of the assumptions above will be retained, but at least the practical bit allocation mechanism will be that of realizable system. Chapter 2 showed that equal quantization stepsize quantizers followed by entropy coding may undergo tractable theoretical evaluation; algorithms based on this technique are the topics of the following chapter.

3.A Statistics of $\triangle R$: “one-shot” estimates

We are interested in deriving the second order statistics of $\Delta R = R - \widehat{R} = R - \underline{x}\underline{x}^T$, and more precisely the $(i,j)th$ block of the $N^2 \times N^2$ matrix $\mathrm{E}\,(vec\Delta R)(vec\Delta R)^T$, where E denotes mathematical expectation. Assuming Gaussian i.i.d. vectors $\underline{x} = [x_1 ... x_N]^T \sim (0, R)$, and writing $R = [\underline{r}_1\ \underline{r}_2 ... \underline{r}_N]$, this block may be written as

$$\begin{aligned} \mathrm{E}\,(vec\Delta R)(vec\Delta R)^T_{block\ i,j} &= E(x_i\underline{x} - \underline{r}_i)(x_j\underline{x} - \underline{r}_j)^T \\ &= \mathrm{E}\,x_i\underline{x}\underline{x}^T x_j - \underline{r}_i\underline{r}_j^T\,. \end{aligned} \tag{3.51}$$

Now let us denote by $\underline{v} = [v_1\ v_2 ... v_N]^T$ white i.i.d. vectors with $v \sim (0, I)$, where I is the identity matrix, and by $R^{\frac{1}{2}}$ the (symmetric) square root of R, $R^{\frac{1}{2}} = [\underline{r}_1^{\frac{1}{2}} \cdots \underline{r}_N^{\frac{1}{2}}]$, where $\underline{r}_j^{\frac{1}{2}}$ denotes the jth column of $R^{\frac{1}{2}}$. Then $\underline{x}$ may be seen as a “colored” version of $\underline{v}$, $\underline{x} = R^{\frac{1}{2}}\underline{v}$, and the first term of eq. (3.51) may then be writen as

$$\begin{aligned} \mathrm{E}\,x_i\underline{x}\underline{x}^T x_j &= E\underline{r}_i^{\frac{T}{2}}\,\underline{v}R^{\frac{1}{2}}\,\underline{v}\underline{v}^T R^{\frac{1}{2}}\,\underline{v}^T\,\underline{r}_j^{\frac{1}{2}} \\ &= \mathrm{E}\,R^{\frac{1}{2}}\,\underbrace{\underline{r}_i^{\frac{T}{2}}\,\underline{v}\underline{v}\underline{v}^T\,\underline{v}^T\,\underline{r}_j^{\frac{1}{2}}}_{A}\,R^{\frac{1}{2}}\,. \end{aligned} \tag{3.52}$$

Let us consider the $(m,n)th$ element $A_{m,n}$ of $A = \underline{r}_i^{\frac{T}{2}}\,\underline{v}\underline{v}\underline{v}^T\,\underline{v}^T\,\underline{r}_j^{\frac{1}{2}}$, which is

$$\begin{aligned} A_{m,n} &= \mathrm{E}\,v_m v_n \left(\sum_k r_{ik}^{\frac{1}{2}} v_k\right)\left(\sum_l r_{jl}^{\frac{1}{2}} v_l\right) \\ &= \underbrace{E\sum_k r_{ik}^{\frac{1}{2}} r_{jk}^{\frac{1}{2}} v_k^2 v_m v_n}_{(a)} + \underbrace{\mathrm{E}\sum_k \sum_{l>k} v_m v_n v_k v_l \left(r_{ik}^{\frac{1}{2}} r_{jl}^{\frac{1}{2}} + r_{il}^{\frac{1}{2}} r_{jk}^{\frac{1}{2}}\right)}_{(b)} \end{aligned} \tag{3.53}$$

where $r_{ps}^{\frac{1}{2}}$ denotes the sth element of $\underline{r}_p$, and the summations run up to N.
We shall now inspect the different cases corresponding to the products involving the $\{v_i\}$.

- Case $m = n$
  - Term (a) : all the terms involved in the summation are nonzero, and two cases should be distinguished
    - * Case $k = m$: the corresponding term yields $\mathrm{E}\,r_{ik}^{\frac{1}{2}} r_{jl}^{\frac{1}{2}} v_k^2 v_m^2 = r_{ik}^{\frac{1}{2}} r_{jl}^{\frac{1}{2}}$
    - * Case $k \neq m$: the corresponding term yields $\mathrm{E}\,r_{ik}^{\frac{1}{2}} r_{jl}^{\frac{1}{2}} v_m^4 = 3r_{ik}^{\frac{1}{2}} r_{jl}^{\frac{1}{2}}$.

    Thus, the term (a) yields then in the case $m = n$

$$\begin{aligned} (a) &= \sum_{k=1}^{N}\left(r_{ik}^{\frac{1}{2}} r_{jk}^{\frac{1}{2}}\right) + 2r_{im}^{\frac{1}{2}} r_{jm}^{\frac{1}{2}} \\ &= \underline{r}_i^{\frac{T}{2}}\,\underline{r}_j^{\frac{1}{2}} + (\ \mathrm{diag}\,\{\underline{r}_j^{\frac{1}{2}}\,\underline{r}_i^{\frac{T}{2}}\})_{m,m} + (\ \mathrm{diag}\,\{\underline{r}_i^{\frac{1}{2}}\,\underline{r}_j^{\frac{T}{2}}\})_{m,m}\,. \end{aligned} \tag{3.54}$$

- Term (b) : all the terms are zero because $l \neq k$ and $\mathrm{E}\,v_m^3 = 0$.

- <u>Case $m \neq n$</u>
  - Term (a) : all terms yield zero by taking expectation for the same reasons as in the previous case
  - Term (b) : two cases may be distinguished :
    - Case $m > n$: Only the term for which $k = n$ and $l = m$ is nonzero, which yields

$$\begin{aligned} \mathrm{E}\, v_m^2 v_n^2 \left(r_{in}^{\frac{1}{2}} r_{jm}^{\frac{1}{2}} + r_{im}^{\frac{1}{2}} r_{jn}^{\frac{1}{2}} \right) &= r_{in}^{\frac{1}{2}} r_{jm}^{\frac{1}{2}} + r_{im}^{\frac{1}{2}} r_{jn}^{\frac{1}{2}} \\ &= \left(\triangleright \{ \underline{r}_j^{\frac{1}{2}} \underline{r}_i^{\frac{T}{2}} \} \right)_{m,n} + \left(\triangleright \{ \underline{r}_i^{\frac{1}{2}} \underline{r}_j^{\frac{T}{2}} \} \right)_{m,n} , \end{aligned} \tag{3.55}$$

where $\triangleright\{.\}$ denotes the strictly lower triangular matrix obtained from the lower triangular part of $\{.\}$, and subscript $_{m,n}$ refers to element m, n.

- Case $m < n$: symmetrically as in the previous case, only the term $k = m$ and $l = n$ is nonzero, which yields

$$r_{in}^{\frac{1}{2}} r_{jm}^{\frac{1}{2}} + r_{im}^{\frac{1}{2}} r_{jn}^{\frac{1}{2}} = \left(\triangleleft \{ \underline{r}_j^{\frac{1}{2}} \underline{r}_i^{\frac{T}{2}} \} \right)_{m,n} + \left(\triangleleft \{ \underline{r}_i^{\frac{1}{2}} \underline{r}_j^{\frac{T}{2}} \} \right)_{m,n} , \tag{3.56}$$

where $\triangleleft\{.\}$ denotes the strictly upper triangular matrix obtained from the upper triangular part of $\{.\}$.

Using (3.54), (3.55) and (3.56), the expectation in eq. (3.52) may now be written as

$$\begin{aligned} \mathrm{E}\, x_i \underline{x}\underline{x}^T x_j &= \mathrm{E}\, R^{\frac{1}{2}} \left(\underline{r}_i^{\frac{T}{2}} \underline{r}_j \tfrac{1}{2} I + \underline{r}_j^{\frac{1}{2}} \underline{r}_i \frac{T}{2} + \underline{r}_i^{\frac{1}{2}} \underline{r}_j \frac{T}{2} \right) R^{\frac{1}{2}} \\ &= r_{ij} R + \underline{r}_j \underline{r}_i^T + \underline{r}_i \underline{r}_j^T . \end{aligned} \tag{3.57}$$

The $(i, j)th$ block in eq. (3.51) is then

$$\begin{aligned} \mathrm{E}\, (vec\Delta R)(vec\Delta R)^T_{block\ i,j} &= r_{ij} R + \underline{r}_j \underline{r}_i^T \\ &= (R \otimes R)_{block\ i,j} + \underline{r}_j \underline{r}_i^T \\ &\approx 2(R \otimes R)_{block\ i,j} , \end{aligned} \tag{3.58}$$

where the approximation is valid for sufficiently highly correlated sources.

3.B Statistics of ΔR: case of K vectors

Let ΔR_i be a particular "one-shot" estimate of $R = \mathrm{E}\,[x_{1,k} ... x_{N,k}]^T [x_{1,k} ... x_{N,k}]$ by means of one vector, $\Delta R_i = R - \underline{x}_i \underline{x}_i^T$, and let $\Delta R^K = R - \frac{1}{K} \sum_{i=1}^{K} \underline{x}_i \underline{x}_i^T = \dfrac{1}{K} \sum_{i=1}^{K} \Delta R_i$ be the estimate of interest. Then

$$\mathrm{E}\, vec\Delta R^K (vec\Delta R^K)^T = \frac{1}{K^2}\, \mathrm{E} \sum_{i=1}^{K} \sum_{j=1}^{K} vec\Delta R_i (vec\Delta R_j)^T , \tag{3.59}$$

where since the ΔR_i are i.i.d., $\mathrm{E}\, vec\Delta R_i (vec\Delta R_j)^T = O_{N^2 \times N^2}$ for $i \neq j$. Thus, using (3.58)

$$\begin{aligned} \mathrm{E}\, vec\Delta R^K (vec\Delta R^K)^T &= \frac{1}{K^2} \underbrace{\sum_{i=1}^{K} \mathrm{E}\, vec\Delta R_i (vec\Delta R_i)^T}_{\approx K\, 2R \otimes R} \\ &\approx \frac{2R\otimes R}{K}. \end{aligned} \tag{3.60}$$

3.C Derivation of (3.37)

In the unitary case, the expected distortion is

$$\mathrm{E}\, \|\underline{\tilde{y}}\|^2_{(\widehat{V},K)} = \mathrm{E} \sum_{i=1}^{N} c2^{-2[r+\frac{1}{2}\log_2 \frac{(\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T)_{ii}}{(\Pi_{i=1}^{N}(\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T)_{ii})^{\frac{1}{N}}}]} (\widehat{V} R_{\underline{xx}} \widehat{V}^T)_{ii}. \tag{3.61}$$

Using the fact $\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T$ is diagonal, we can write (3.61) as

$$\mathrm{E}\, \|\underline{\tilde{y}}\|^2_{(\widehat{V},K)} = \mathrm{E}\, c2^{-2r} \left(\det \widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T\right)^{\frac{1}{N}} \sum_{i=1}^{N} \frac{(\widehat{V} R_{\underline{xx}} \widehat{V}^T)_{ii}}{(\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T)_{ii}}. \tag{3.62}$$

Because of the unimodularity of $\widehat{V}$, the determinant in (3.62) may be written as

$$\begin{aligned} \det \widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T &= \det(R_{\underline{xx}} + \Delta R) \\ &= (\det R_{\underline{xx}}) \det(I + R_{\underline{xx}}^{-1} \Delta R). \end{aligned} \tag{3.63}$$

The sum in (3.62) may be written as

$$\begin{aligned} \sum_{i=1}^{N} \frac{(\widehat{V} R_{\underline{xx}} \widehat{V}^T)_{ii}}{(\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T)_{ii}} &= \mathrm{tr}\,\{(\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T)^{-1/2}\widehat{V} R_{\underline{xx}} \widehat{V}^T (\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T)^{-1/2}\} \\ &= tr\{\widehat{V} R_{\underline{xx}} \widehat{V}^T (\widehat{V}\widehat{R}_{\underline{xx}}\widehat{V}^T)^{-1}\} = \mathrm{tr}\,\{\widehat{V} R_{\underline{xx}} \widehat{R}_{\underline{xx}}^{-1} \widehat{V}^{-1}\} \\ &= \mathrm{tr}\,\{R_{\underline{xx}} \widehat{R}_{\underline{xx}}^{-1})\} = tr\{R_{\underline{xx}}(R_{\underline{xx}} + \Delta R)^{-1}\} \\ &= \mathrm{tr}\,\{(I + R_{\underline{xx}}^{-1}\Delta R)^{-1}\}. \end{aligned} \tag{3.64}$$

Thus, (3.62) is equivalent to

$$\mathrm{E}\, \|\underline{\tilde{y}}\|^2_{(\widehat{V},K)} = \mathrm{E}\, \|\underline{\tilde{y}}\|^2_{V} \left(\frac{1}{N}\ \mathrm{E}\ (\det(I + R_{\underline{xx}}^{-1}\Delta R))^{\frac{1}{N}}\, \mathrm{tr}\,\{\left(I + R_{\underline{xx}}^{-1}\Delta R\right)^{-1}\}\right), \tag{3.65}$$

which also the distortion obtained in the causal case.

In order to compute the expectation of $\mathrm{E}\, \|\underline{\tilde{y}}\|^2_{(\widehat{V},K)}$, let us develop (3.61) as

$$\mathrm{E}\, \|\underline{\tilde{y}}\|^2_{(\widehat{V},K)} = \mathrm{E}\, c2^{-2r} \left(\det\ \mathrm{diag}\,\{\widehat{V} R_{\underline{xx}} \widehat{V}^T\}\right)^{\frac{1}{N}} \left(\prod_{i=1}^{N}(1 + \frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{xx}} \widehat{V}^T)_{ii}})\right)^{\frac{1}{N}} \sum_{i=1}^{N} \left(1 + \frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{xx}} \widehat{V}^T)_{ii}}\right)^{-1}. \tag{3.66}$$

The determinant in (3.66) may also be written as

$$\begin{aligned} \det \text{ diag}\{\widehat{V} R_{\underline{xx}} \widehat{V}^T\} &= \prod_{i=1}^{N} (\lambda_i + \delta_{\lambda_i}) \\ &= (\prod_{i=1}^{N} \lambda_i)(\prod_{i=1}^{N} 1 + \frac{\delta_{\lambda_i}}{\lambda_i}) \\ &\approx (\prod_{i=1}^{N} \lambda_i)(1 + \sum_{i=1}^{N} \frac{\delta_{\lambda_i}}{\lambda_i}), \end{aligned} \tag{3.67}$$

where λ_i ans δ_{λ_i} are the diagonal elements of Λ and $\Delta\Lambda$, which is defined by

$$\widehat{V} R_{\underline{xx}} \widehat{V}^T = \Lambda + \Delta\Lambda. \tag{3.68}$$

Now (3.66) may then be approximated as

$$\begin{aligned} &\text{E}\,||\underline{\tilde{y}}||^2_{(\widehat{V},K)} = \text{E}\,||\underline{\tilde{y}}||^2_{(K)} \left(1 + \tfrac{1}{N} \sum_{i=1}^{N} \frac{\delta_{\lambda_i}}{\lambda_i}\right) \left(\prod_{i=1}^{N} (1 + \frac{(\widehat{V} \Delta R \widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{xx}} \widehat{V}^T)_{ii}})\right)^{\frac{1}{N}} \sum_{i=1}^{N} \left(1 + \frac{(\widehat{V} \Delta R \widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{xx}} \widehat{V}^T)_{ii}}\right)^{-1} \\ &\approx \text{E}\,||\underline{\tilde{y}}||^2_{(K)} \\ &\left(1 + \; E\tfrac{1}{N} \sum_{i=1}^{N} \frac{\delta_{\lambda_i}}{\lambda_i} + \; \text{E} \; \frac{N-1}{2N^2} \sum_{i=1}^{N} (\frac{(\widehat{V} \Delta R \widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{xx}} \widehat{V}^T)_{ii}})^2 - \; \text{E} \; \frac{1}{N^2} \sum_{i} \sum_{j>i} \frac{(\widehat{V} \Delta R \widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{xx}} \widehat{V}^T)_{ii}} \frac{(\widehat{V} \Delta R \widehat{V}^T)_{jj}}{(\widehat{V} R_{\underline{xx}} \widehat{V}^T)_{jj}}\right). \end{aligned} \tag{3.69}$$

Computation of the second expectation in (3.69).

Using the unitarity of $\widehat{V} = V + \Delta V$, we have

$$\text{diag}\,\{\Delta\Lambda\} = diag\{\Delta V^T R_{\underline{xx}} \Delta V - \Delta V^T \Delta V \Lambda\}. \tag{3.70}$$

The expectation of the diagonal elements of the first term in (3.70) is

$$\text{E} \; (\Delta V^T R_{\underline{xx}} \Delta V)_{ii} = \text{tr} \; \text{E} \; R_{\underline{xx}} \Delta V_i \Delta V_i^T = \frac{\lambda_i}{K} \sum_{j \neq i} \frac{\lambda_j^2}{(\lambda_j - \lambda_i)^2}, \tag{3.71}$$

where we have used the following classical result in perturbation theory of matrices [77], for sufficiently high K

$$\text{E} \; \Delta V_i \Delta V_i^T = \frac{\lambda_i}{K} \sum_{j \neq i} \frac{\lambda_j}{(\lambda_j - \lambda_i)^2} T_j T_j^T. \tag{3.72}$$

The expectation of the diagonal elements of the second term in (3.70) is

$$\text{E} \; (\Delta V^T \Delta V \Lambda)_{ii} = \lambda_i \, \text{E}\,(\Delta V_i^T \Delta V_i) = \frac{\lambda_i^2}{K} \sum_{j \neq i} \frac{\lambda_j}{(\lambda_j - \lambda_i)^2}. \tag{3.73}$$

Hence, we get from (3.70,3.71), and (3.73)

$$\text{E}\,\delta\lambda_i = \frac{\lambda_i}{K} \sum_{j \neq i} \frac{\lambda_j^2 - \lambda_i \lambda_j}{(\lambda_j - \lambda_i)^2} = \frac{\lambda_i}{K} \sum_{j \neq i} \frac{\lambda_j}{\lambda_j - \lambda_i}, \tag{3.74}$$

from which it is easy to show that

$$\mathrm{E}\,\sum_{i=1}^{N}\frac{\delta\lambda_i}{\lambda_i}=\frac{1}{K}\sum_{i=1}^{N}\sum_{j\neq i}\frac{\lambda_j}{\lambda_j-\lambda_i}=\frac{1}{K}\frac{N(N-1)}{2}. \tag{3.75}$$

Computation of the third and fourth expectations in (3.69).

The perturbing term $\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V}R_{\underline{x}\underline{x}}\widehat{V}^T)_{ii}}$ may also be approximated as

$$\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V}R_{\underline{x}\underline{x}}\widehat{V}^T)_{ii}}=\frac{V\Delta RV^T+V\Delta R\Delta V^T+\Delta V\Delta RV^T+\Delta V\Delta R\Delta V^T}{VR_{\underline{x}\underline{x}}V^T+VR_{\underline{x}\underline{x}}\Delta V^T+\Delta VR_{\underline{x}\underline{x}}V^T+\Delta VR_{\underline{x}\underline{x}}\Delta V^T}\approx\frac{V\Delta RV^T}{VR_{\underline{x}\underline{x}}V^T} \tag{3.76}$$

Let $\widehat{\Lambda}$ be

$$\begin{aligned}\widehat{\Lambda} &= V\widehat{R}_{\underline{x}\underline{x}}V^T=VR_{\underline{x}\underline{x}}V^T+V\Delta RV^T\\ &= \tfrac{1}{K}\sum_{i=1}^{N}V\underline{x}_i\underline{x}_i^TV^T\\ &= \tfrac{1}{K}\sum_{i=1}^{N}\underline{y}_i\underline{y}_i^T\,.\end{aligned} \tag{3.77}$$

Thus, $V\Delta RV^T=\Lambda-\frac{1}{K}\sum_{i=1}^{N}\underline{y}_i\underline{y}_i^T$: $(V\Delta RV^T)_{ii}$ is a real zero mean Gaussian random variable, corresponding to the estimation error of Λ obtained with a covariance matrix computed with K vectors. Hence, we have $\mathrm{E}\,vec(V\Delta RV^T)(vecV\Delta RV^T)^T\approx\frac{2\Lambda\otimes\Lambda}{K}$, whence

$$\sum_{i=1}^{N}\left(\frac{(V\Delta RV^T)_{ii}}{(VR_{\underline{x}\underline{x}}V^T)_{ii}}\right)^2=\frac{2N}{K}, \tag{3.78}$$

and

$$\sum_{i}\sum_{j>i}\frac{(V\Delta RV^T)_{ii}}{(VR_{\underline{x}\underline{x}}V^T)_{ii}}\frac{(V\Delta RV^T)_{jj}}{(VR_{\underline{x}\underline{x}}V^T)_{jj}}=\sum_{i}\sum_{j>i}\frac{2}{K}\frac{\Lambda_{ij}^2}{\lambda_i\lambda_j}=0. \tag{3.79}$$

Finally, the expected distortion for the KLT when the transformation is based on K vectors is, under high resolution assumption

$$\mathrm{E}\,||\underline{\tilde{y}}||^2_{(\widehat{V},K)}\approx\mathrm{E}\,||\underline{\tilde{y}}||^2_{V}\left(1+\frac{N-1}{K}\left[\frac{1}{2}+\frac{1}{N}\right]\right). \tag{3.80}$$

3.D Derivation of (3.44)

The first term of eq.(3.43) may be written as

$$\begin{aligned}&c2^{-2[r+\frac{1}{2}log_2\frac{(R_{\underline{x}^q\underline{x}^q})_{ii}}{(\Pi_{i=1}^{N}(R_{\underline{x}^q\underline{x}^q})_{ii})^{\frac{1}{N}}}]}(R_{\underline{x}^q\underline{x}^q})_{ii}\,\mathrm{E}\left(\prod_{i=1}^{N}1+\frac{(\Delta R)_{ii}}{(R_{\underline{x}^q\underline{x}^q})_{ii}}\right)^{\frac{1}{N}}\sum_{i=1}^{N}\left(1+\frac{(\Delta R)_{ii}}{(R_{\underline{x}^q\underline{x}^q})_{ii}}\right)^{-1}\\ &\approx\ \mathrm{E}\,||\underline{\tilde{y}}||^2_I c2^{-2r}\left(\det(\,\mathrm{diag}\,\{R_{\underline{x}^q\underline{x}^q}\})\right)^{\frac{1}{N}}\\ &\quad\times\left(1+\ \mathrm{E}\ \tfrac{N-1}{2N^2}\sum_{i=1}^{N}(\frac{(\Delta R)_{ii}}{(R_{\underline{x}^q\underline{x}^q})_{ii}})^2-\ \mathrm{E}\ \frac{1}{N^2}\sum_{i}\sum_{j>i}\frac{(\Delta R)_{ii}}{(R_{\underline{x}^q\underline{x}^q})_{ii}}\frac{(\Delta R)_{jj}}{(R_{\underline{x}^q\underline{x}^q})_{jj}}\right).\end{aligned} \tag{3.81}$$

The equality concerning the determinant comes from a factorization similar to (3.5). The expectations in (3.81) are computed in the same manner as in section 3.2. Note however that in this case, the r.v. ΔR corresponds to the estimation error of $R_{\underline{x}\underline{x}}$, which is not the covariance matrix of Gaussian r.v.s because of the uniformly distributed quantization noise q_i perturbing the x_i. Since this perturbation is small we assume that ΔR can be considered as a zero mean r.v. with covariance matrix

$$\mathrm{E}\, vec(\Delta R)\,(vec(\Delta R))^T \approx \frac{2}{K} R_{\underline{x}^q\underline{x}^q} \otimes R_{\underline{x}^q\underline{x}^q}. \tag{3.82}$$

With (3.82), the second expectation in (3.81) may be approximated as

$$\mathrm{E}\ \frac{N-1}{2N^2}\sum_{i=1}^{N}\left(\frac{(\Delta R)_{ii}}{(R_{\underline{x}^q\underline{x}^q})_{ii}}\right)^2 \approx \frac{N-1}{2N^2}\sum_{i=1}^{N}\frac{2(R_{\underline{x}^q\underline{x}^q})_{ii}^2}{K(R_{\underline{x}^q\underline{x}^q})_{ii}^2} = \frac{N-1}{2N^2}\frac{2N}{K} = \frac{N-1}{NK}, \tag{3.83}$$

and the third expectation as

$$\begin{aligned}\mathrm{E}\ \tfrac{1}{N^2}\sum_i\sum_{j>i}\frac{(\Delta R)_{ii}}{(R_{\underline{x}^q\underline{x}^q})_{ii}}\frac{(\Delta R)_{jj}}{(R_{\underline{x}^q\underline{x}^q})_{jj}} \quad &\approx \quad \tfrac{2}{K}\sum_i\sum_{j>i}\frac{(R_{\underline{x}^q\underline{x}^q})_{ij}^2}{(R_{\underline{x}^q\underline{x}^q})_{ii}(R_{\underline{x}^q\underline{x}^q})_{jj}} \\ &\approx \quad \tfrac{2}{K}\|\triangleright\left((\,\mathrm{diag}\,\{R_{\underline{x}^q\underline{x}^q}\})^{1/2}R_{\underline{x}^q\underline{x}^q}(\,\mathrm{diag}\,\{R_{\underline{x}^q\underline{x}^q}\})^{1/2}\right)\|_F^2.\end{aligned} \tag{3.84}$$

If D^q denotes $\mathrm{diag}\,\{R_{\underline{x}^q\underline{x}^q}\}$, we obtain

$$\begin{aligned}\mathrm{E}\ \tfrac{1}{N^2}\sum_i\sum_{j>i}\frac{(\Delta R)_{ii}}{(R_{\underline{x}^q\underline{x}^q})_{ii}}\frac{(\Delta R)_{jj}}{(R_{\underline{x}^q\underline{x}^q})_{jj}} \quad &\approx \quad \tfrac{1}{K}\left(\|(D^q)^{-\frac{1}{2}}R_{\underline{x}\underline{x}}(D^q)^{-\frac{1}{2}}\|_F^2 - \|\,\mathrm{diag}\,\{(D^q)^{-\frac{1}{2}}R_{\underline{x}\underline{x}}(D^q)^{-\frac{1}{2}}\}\|_F^2\right) \\ &\approx \quad \tfrac{1}{K}(\,\mathrm{tr}\,\{R_{\underline{x}\underline{x}}(D^q)^{-1}R_{\underline{x}\underline{x}}(D^q)^{-1}\}).\end{aligned} \tag{3.85}$$

The second term in (3.43) is small because of σ_q^2, and we neglect the estimation errors in this term (estimation errors being itself small for sufficiently high K), so that we make the approximation

$$\sigma_q^2 E\sum_{i=1}^{N} c2^{-2[r+\frac{1}{2}log_2\frac{(\hat{R}_{\underline{x}^q\underline{x}^q})_{ii}}{(\Pi_{i=1}^{N}(\hat{R}_{\underline{x}^q\underline{x}^q})_{ii})^{\frac{1}{N}}}]} \approx \sigma_q^2 c2^{-2r}\left(\det(\,\mathrm{diag}\,\{R_{\underline{x}^q\underline{x}^q}\})\right)^{\frac{1}{N}}\left[-\sum_{i=1}^{N}\frac{1}{(R_{\underline{x}^q\underline{x}^q})_{ii}}\right]. \tag{3.86}$$

Finally, using (3.81), (3.85) and (3.86), the expected distortion for Identity transform with quantization and estimation noise may, for sufficiently high resolution and K, be written as

$$\begin{aligned}\mathrm{E}\,\|\tilde{\underline{y}}\|_{(I,K,q)}^2 \approx\ &\mathrm{E}\,\|\tilde{\underline{y}}\|_I^2\left(\det(I+\sigma_q^2(\,\mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{-1})\right)^{1/N} \\ &\times\left[1+\tfrac{1}{K}\left[1-\frac{\mathrm{tr}\,\{R_{\underline{x}^q\underline{x}^q}D^{q-1}R_{\underline{x}^q\underline{x}^q}D^{q-1}\}}{N^2}\right]-\frac{\sigma_q^2}{N}\,\mathrm{tr}\,\{(\,\mathrm{diag}\,R_{\underline{x}^q\underline{x}^q})^{-1}\}\right].\end{aligned} \tag{3.87}$$

3.E Derivation of (3.46)

By writing $\widehat{V} R_{\underline{x}\underline{x}} \widehat{V}^T = \widehat{V} R_{\underline{x}^q\underline{x}^q} \widehat{V}^T - \sigma_q^2 \widehat{V}\widehat{V}^T = \widehat{V} R_{\underline{x}^q\underline{x}^q} \widehat{V}^T - \sigma_q^2 I$ in (3.45), we get

$$\begin{aligned}
\mathrm{E}\,||\underline{\tilde{y}}||^2_{(\widehat{V},K,q)} &= \mathrm{E}\sum_{i=1}^{N} c2^{-2[r+\frac{1}{2}log_2\frac{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}{(\Pi_{i=1}^N(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii})^{\frac{1}{N}}}]}\left(\prod_{i=1}^{N}(1+\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}})\right)^{\frac{1}{N}} \\
&\quad\times\left(\frac{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}-\frac{\sigma_q^2}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}\right)\left(1+\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}\right) \\
&= \mathrm{E}\,c2^{-2r}\left(\det\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T\right)^{\frac{1}{N}}\left(\prod_{i=1}^{N}(1+\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}})\right)^{\frac{1}{N}} \\
&\quad\times\left[\sum_{i=1}^{N}\left(1+\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}\right)^{-1}-\sigma_q^2\sum_{i=1}^{N}((\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii})^{-1}\right]
\end{aligned} \tag{3.88}$$

Rewriting the last equality of eq.(3.88), we have

$$\begin{aligned}
\mathrm{E}\,||\underline{\tilde{y}}||^2_{(\widehat{V},K,q)} &= \mathrm{E}\,c2^{-2r}\left(\det\ \mathrm{diag}\{\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T\}\right)^{\frac{1}{N}}\left(\prod_{i=1}^{N}(1+\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}})\right)^{\frac{1}{N}} \\
&\quad\times\left[\sum_{i=1}^{N}\left(1+\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}\right)^{-1}-\sigma_q^2\sum_{i=1}^{N}\left((\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}\right)^{-1}\right].
\end{aligned} \tag{3.89}$$

Now, let $\widehat{\Lambda}^q = \widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T = \Lambda^q + \Delta\Lambda^q = \Lambda + \sigma_q^2 I + \Delta\Lambda^q$, and let $\delta^q_{\lambda_i}$ be the diagonal elements of $\Delta\Lambda^q$. Then, the first term of (3.89) may be approximated as

$$\begin{aligned}
&\mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{V}}\left(1+\tfrac{1}{N}\sum_{i=1}^{N}\frac{\delta^q_{\lambda_i}}{\Lambda^q_{ii}}\right)\left(\prod_{i=1}^{N}(1+\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}})\right)^{\frac{1}{N}}\sum_{i=1}^{N}\left(1+\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}\right)^{-1} \\
&\approx\ \mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{V}} \\
&\left(1+\mathrm{E}\,\tfrac{1}{N}\sum_{i=1}^{N}\frac{\delta^q_{\lambda_i}}{\Lambda^q_{ii}}+\mathrm{E}\,\frac{N-1}{2N^2}\sum_{i=1}^{N}\left(\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}\right)^2-\mathrm{E}\,\frac{1}{N^2}\sum_{i}\sum_{j>i}\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}}\frac{(\widehat{V}\Delta R\widehat{V}^T)_{jj}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{jj}}\right).
\end{aligned} \tag{3.90}$$

Using an analysis similar to (3.69) and using the same classical result in pertubation theory of matrices as in the previous section [77], one can show that

$$\mathrm{E}\sum_{i=1}^{N}\frac{\delta\lambda_i^q}{\Lambda^q_{ii}} = \frac{1}{K}\sum_{i=1}^{N}\sum_{j\neq i}\frac{\lambda_j^q}{\lambda_j^q-\lambda_i^q} = \frac{1}{K}\frac{N(N-1)}{2}. \tag{3.91}$$

Also, the expectation of the second term in (3.90) may be computed as in (3.69). By using the first order approximation

$$\frac{(\widehat{V}\Delta R\widehat{V}^T)_{ii}}{(\widehat{V} R_{\underline{x}^q\underline{x}^q}\widehat{V}^T)_{ii}} \approx \frac{(V\Delta R V^T)_{ii}}{(V R_{\underline{x}^q\underline{x}^q} V^T)_{ii}}, \tag{3.92}$$

and by writing $V\Delta R V^T$ as $\Lambda^q - \frac{1}{K}\sum_{i=1}^{N} Y_i^q Y_i^{qT}$, the random variable $(V\Delta R V^T)_{ii}$ corresponds now to the estimation error of Λ^q obtained with a covariance matrix computed with K quantized vectors. As in (3.82) however, this is again an approximation since the y_i^q are not Gaussian. Thus we assume $\mathrm{E}\, vec(V\Delta R V^T)(vec(V\Delta R V^T))^T \approx \frac{2\Lambda^q \otimes \Lambda^q}{K}$, whence

$$\sum_{i=1}^{N} \left(\frac{(V\Delta R V^T)_{ii}}{(V R_{\underline{x}\underline{x}} V^T)_{ii}}\right)^2 \approx \frac{2N}{K}, \tag{3.93}$$

and

$$\sum_i \sum_{j>i} \frac{(V\Delta R V^T)_{ii}}{(V R_{\underline{x}^q\underline{x}^q} V^T)_{ii}} \frac{(V\Delta R V^T)_{jj}}{(V R_{\underline{x}^q\underline{x}^q} V^T)_{jj}} = \sum_i \sum_{j>i} \frac{2}{K} \frac{(\Lambda^q)^2_{ij}}{\lambda_i^q \lambda_j^q} \approx 0. \tag{3.94}$$

Thus, using the unimodularity of $\widehat{T}$, the first term becomes

$$\begin{array}{l} c2^{-2r}(\det R_{\underline{x}^q\underline{x}^q})^{1/N} \left[1 + \frac{N(N-1)}{2K} + \frac{2N(N-1)}{2N^2K}\right] \\ \approx c2^{-2r}(\det R_{XX})^{1/N} \left(\det(I + \sigma_q^2 (R_{XX})^{-1})\right)^{1/N} \left[1 + \frac{1}{K}(\frac{N-1}{2} + \frac{N-1}{N})\right]. \end{array} \tag{3.95}$$

The second term in (3.88) may approximated under the assumptions of high resolution and high K as

$$\begin{aligned} \sigma_q^2\, \mathrm{E}\, c^{-2r} (\det R_{\underline{x}^q\underline{x}^q})^{\frac{1}{N}} &\approx \sum_{i=1}^{N} \frac{1}{(V R_{\underline{x}^q\underline{x}^q} V^T)_{ii}} \\ &\approx \sigma_q^2\, \mathrm{E}\, c^{-2r} (\det R_{\underline{x}^q\underline{x}^q})^{\frac{1}{N}}\, \mathrm{tr}\{(\Lambda^q)^{-1}\} \\ &\approx \sigma_q^2 c2^{-2r} (\det R_{XX})^{1/N} \left(\det(I + \sigma_q^2 (R_{XX})^{-1})\right)^{1/N}\, \mathrm{tr}\{(\Lambda^q)^{-1}\}. \end{aligned} \tag{3.96}$$

Finally, the expected distortion for the KLT when the transformation is based on K quantized vectors is for high K and under high resolution assumption

$$\mathrm{E}\, ||\underline{\tilde{y}}||^2_{(\widehat{V},K,q)} \approx \mathrm{E}\, ||\underline{\tilde{y}}||^2_{(K)} \left(\det(I + \sigma_q^2 (R_{XX})^{-1})\right)^{1/N} \times \left[1 + \frac{N-1}{K}\left[\frac{1}{2} + \frac{1}{N}\right] - \frac{\sigma_q^2}{N}\, \mathrm{tr}\{(\Lambda^q)^{-1}\}\right]. \tag{3.97}$$

3.F Derivation of (3.49)

The expression (3.48) can be developped as

$$\begin{aligned}
\mathrm{E}\,||\underline{\tilde{y}}||^2_{(\widehat{L}',K,q)} &= \mathrm{E}\sum_{i=1}^{N} c2^{-2\left[r+\frac{1}{2}log_2\frac{(\widehat{L}'R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}}{\left(\prod_{i=1}^{N}(\widehat{L}'R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}\right)^{\frac{1}{N}}}\right]}\left(\prod_{i=1}^{N}(1+\frac{(\widehat{L}'\Delta R\widehat{L}'^T)_{ii}}{(\widehat{L}R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}})\right)^{\frac{1}{N}} \\
&\quad\times\left(\frac{(\widehat{L}'R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}}{(\widehat{L}'R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}}-\frac{\sigma_q^2}{(\widehat{L}'R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}}\right)\left(1+\frac{(\widehat{L}'\Delta R\widehat{L}'^T)_{ii}}{(\widehat{L}R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}}\right) \\
&= \mathrm{E}\,c2^{-2r}\left(\det\ \mathrm{diag}\{\widehat{L}'R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T\}\right)^{\frac{1}{N}}\left(\prod_{i=1}^{N}(1+\frac{(\widehat{L}'\Delta R\widehat{L}'^T)_{ii}}{(\widehat{L}R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}})\right)^{\frac{1}{N}} \\
&\quad\times\left[\sum_{i=1}^{N}\left(1+\frac{(\widehat{L}'\Delta R\widehat{L}'^T)_{ii}}{(\widehat{L}R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}}\right)^{-1}-\sigma_q^2\sum_{i=1}^{N}((\widehat{L}'R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii})^{-1}\right].
\end{aligned} \tag{3.98}$$

Now, let $\widehat{R}^q_{\underline{yy}}$ be $\widehat{L}'R_{\underline{x}^q\underline{x}^q}\widehat{L}' = R^q_{\underline{yy}} + \Delta R^q_{\underline{yy}}$, where $R^q_{\underline{yy}} = L'R_{\underline{x}^q\underline{x}^q}L'^T$, and let δ_{y^q} be the diagonal elements of $\Delta R^q_{\underline{yy}}$. Then, the first term of (3.98) may be written as

$$\begin{aligned}
&\mathrm{E}\,||\underline{\tilde{y}}||^2_{(\widehat{L}',K,q)} = \mathrm{E}\,||\underline{\tilde{y}}||^2_L\left(1+\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_{y^q}}{(R^q_{\underline{yy}})_{ii}}\right)\left(\prod_{i=1}^{N}(1+\frac{(\widehat{L}'\Delta R\widehat{L}'^T)_{ii}}{(\widehat{L}R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}})\right)^{\frac{1}{N}}\sum_{i=1}^{N}\left(1+\frac{(\widehat{L}'\Delta R\widehat{L}'^T)_{ii}}{(\widehat{L}R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}}\right)^{-1} \\
&\approx E||\underline{\tilde{y}}||^2_L \\
&\times\left(1+E\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_{y^q}}{(R^q_{\underline{yy}})_{ii}}+\mathrm{E}\,\frac{N-1}{2N^2}\sum_{i=1}^{N}(\frac{(\widehat{L}'\Delta R\widehat{L}'^T)_{ii}}{(\widehat{L}R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}})^2-\mathrm{E}\,\frac{1}{N^2}\sum_i\sum_{j>i}\frac{(\widehat{L}'\Delta R\widehat{L}'^T)_{ii}}{(\widehat{L}R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}}\frac{(\widehat{L}'\Delta R\widehat{L}'^T)_{jj}}{(\widehat{L}'R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{jj}}\right)
\end{aligned} \tag{3.99}$$

Using a similar analysis as in (3.69), one can show that

$$\mathrm{E}\sum_{i=1}^{N}\frac{\delta_{y^q}}{(R^q_{\underline{yy}})_{ii}} = \frac{1}{K}\sum_{i=1}^{N}\sum_{j\neq i}\frac{(R^q_{\underline{yy}})_{jj}}{(R^q_{\underline{yy}})_{jj}-(R^q_{\underline{yy}})_{ii}} = \frac{1}{K}\frac{N(N-1)}{2}. \tag{3.100}$$

Thus, using the unimodularity of $\widehat{L}'$, the first term may be approximated as

$$\begin{aligned}
c2^{-2r}(\det R_{\underline{x}^q\underline{x}^q})^{1/N}\left[1+\frac{N(N-1)}{2K}+\frac{2N(N-1)}{2N^2K}\right] &= c2^{-2r}(\det R_{\underline{xx}})^{1/N} \\
\times\left(\det(I+\sigma_q^2(R_{\underline{xx}})^{-1})\right)^{1/N}&\left[1+\frac{1}{K}\left(\frac{N-1}{2}+\frac{N-1}{K}\right)\right].
\end{aligned} \tag{3.101}$$

The expectation of the second term in (3.98) can be computed as in (3.69) also. By using the approximation

$$\frac{(\widehat{L}'\Delta R\widehat{L}'^T)_{ii}}{(\widehat{L}R_{\underline{x}^q\underline{x}^q}\widehat{L}'^T)_{ii}} \approx \frac{(L'\Delta RL'^T)_{ii}}{(L'R_{\underline{x}^q\underline{x}^q}L'^T)_{ii}} \tag{3.102}$$

We can write $L'\Delta RL'^T = R^q_{\underline{yy}} - \frac{1}{K}\sum_{i=1}^{N}\underline{y}^q_i\underline{y}^{qT}_i$: the random variable $(L'\Delta RL'^T)_{ii}$ corresponds now to the estimation error of $R^q_{\underline{yy}}$ obtained with a covariance matrix computed with K quantized vectors. Again, we make the approximation of Gaussianity for y_i^q. Thus we assume $\mathrm{E}\, vec(L'\Delta R L^{'T})(vec(L'\Delta R L^{'T}))^T \approx \frac{2R^q_{\underline{yy}} \otimes R^q_{\underline{yy}}}{K}$, whence

$$\sum_{i=1}^{N} \left(\frac{(L'\Delta R L^{'T})_{ii}}{(L' R_{\underline{xx}} L^{'T})_{ii}} \right)^2 \approx \frac{2N}{K}, \tag{3.103}$$

and

$$\sum_i \sum_{j>i} \frac{(L'\Delta R L^{'T})_{ii}}{(L' R_{\underline{x}^q\underline{x}^q} L^{'T})_{ii}} \frac{(L'\Delta R L^{'T})_{jj}}{(L' R_{\underline{x}^q\underline{x}^q} L^{'T})_{jj}} = \sum_i \sum_{j>i} \frac{2}{K} \frac{(R^q_{\underline{yy}})^2_{ij}}{(R^q_{\underline{yy}})_{ii} (R^q_{\underline{yy}})_j} \approx 0. \tag{3.104}$$

The second term in (3.98) may approximated under the assumptions of high resolution and high K as

$$\begin{aligned} \sigma_q^2 \,\mathrm{E}\, c^{-2r} (\det R_{\underline{x}^q\underline{x}^q})^{\frac{1}{N}} &\approx \sum_{i=1}^{N} \frac{1}{(L' R_{\underline{x}^q\underline{x}^q} L^{'T})_{ii}} \\ &\approx \sigma_q^2 \,\mathrm{E}\, c^{-2r} (\det R_{\underline{x}^q\underline{x}^q})^{\frac{1}{N}} \operatorname{tr}\{(R^q_{\underline{yy}})^{-1}\} \\ &\approx \sigma_q^2 c 2^{-2r} (\det R_{\underline{xx}})^{1/N} \left(\det(I + \sigma_q^2 (R_{\underline{xx}})^{-1})\right)^{1/N} \operatorname{tr}\{(R'_{\underline{yy}})^{-1}\} \end{aligned} \tag{3.105}$$

Finally, using the obtained expressions for the first and second terms of (3.98), the expected distortion for the LDU when the transformation is based on K quantized vectors is for high K and under high resolution assumption

$$\mathrm{E}\, ||\underline{\tilde{y}}||^2_{(\widehat{L}',K,q)} \approx \mathrm{E}\, ||\underline{\tilde{y}}||^2_{L} \left(\det(I + \sigma_q^2 (R_{\underline{xx}})^{-1})\right)^{1/N} \left[1 + \frac{N-1}{K} \left[\frac{1}{2} + \frac{1}{N} \right] - \frac{\sigma_q^2}{N} \operatorname{tr}\{(R'_{\underline{yy}})^{-1}\} \right]. \tag{3.106}$$

Chapter 4

Rate-Distortion Analysis of Practical Backward Adaptive Transform Coding Schemes

The main advantage of backward over forward adaptive coding schemes is to update the coding parameters with the data available at the decoder, avoiding thereby any excess bitrate. The algorithms presented in this chapter aims of evaluating the performance of practical backward adaptive transform coding schemes. Their performance are analyzed in terms of rate and distortion, for the causal transform introduced in chapter 2, and for the Karhunen-Loève tranform. The optimal bit allocation rule, which somewhat limits, from a practical viewpoint, the results of the previous chapter, is replaced here by a simple (equal stepsize) quantization rule. In a first step, algorithms with constant stepsizes are considered: only the tranforms are backward adaptive. In a second step, both the stepsize and the transforms rely on backward adaptation. In this framework, the transform coding system is designed a priori to operate at a particular (target) rate-distortion point. The question is to know whether the proposed algorithms will converge or not to this point. For two algorithms, we evaluate the resulting expected distortion w.r.t. the number of vectors available at the decoder, as the distortion to which the systems converge. The rate is then measured by the 0*th order entropy of the corresponding sequence of quantized data (asymptotically in the data length). A high resolution analysis shows that for an algorithm using Sheppard's correction on the second order moment estimates, the performance of the system should converge to the target rate-distortion point. Without this correction, the effects of backward adaptation tend to move the operational rate-distortion point of the system from the target point by the same term for both transforms.*

4.1 Introduction

For non- or locally- stationary data, the efficiency of transform coding relies on the updating of the coding parameters according to the source statistics changes. These updates aim to keep the performance of the structure close to a predetermined rate-distortion trade-off. Classically, they are sent as side information to the decoder, though this excess bitrate could be saved by using closed-loop, or backward adaptive algorithms.

A first contribution of this work is a numerical evaluation of practical algorithms using equal and constant (w.r.t. time) quantization stepsizes. Choosing a stepsize is equivalent to choose a target point of the rate-distortion function of the source. If the source statictics do not change, the system may converge to the target point, assuming that the transforms will converge to the optimal transforms (that is, designed with *a priori* knowledge of the source). We will not try to prove convergence results in this first part, nor in the rest of the paper. Instead, empirical evidence, or analytical evaluation based on small perturbations will be proposed.

The first results for constant stepsize will provide empirical evidence that these systems converge. These results are complementary to those established in [54], which regard the universality of the KLT in this framework; also, the results of section 4.3 suggest the universality of the LDU at high rates. The sense given to *universality* in this work is that of [54]: the ability of an adaptive system to provide, asymptotically in the data length, the optimal rate-distortion performance for a given class of source, which in our case is Gaussian. The drawback of using a constant is however that the rate may unacceptably vary in the case where the statistics of the source change. The distortion is fixed, but may become unacceptable as well w.r.t. to the energy of the source. As in adaptive predictive quantization, one may prefer algorithms relying on updating not only the transforms, but also the quantization stepsizes.

We propose therefore in a second step to model the effects of the backward adaptation for two simple algorithms using adaptive stepsizes, and for two different transforms, the unitary KLT and the causal LDU transform. A transform coding scheme is in the ideal case designed to reach a target point of the rate-distortion function of the source. This point (r_0, D_0,R) is characterized by the covariance matrix R, a target rate r_0 and the corresponding target distortion D_0. Assuming now that we use some backward adapted algorithms, an interesting question is to know if the coding performance will converge or not to the target rate-distortion point, and if yes, how fast.

The theoretical comparison between causal and unitary approaches presented in the last chapter did not describe how practical backward adaptive transform algorithms would perform. This is the aim of the present work, where we propose an analysis based on small perturbations. The investigated methods correspond to actual, implementable algorithms, but again, we have to make several assumptions. We suppose that the coding structure deals with a (possibly locally) stationary Gaussian source with covariance matrix R[1], whose vector samples are independent and identically distributed. The results regarding the rate are asymp-

[1] In order to simplify the notations, R will denote the covariance matrix $R_{\underline{x}\underline{x}}$ in this chapter.

totic in the data length. Moreover, we assume that the entropy coder possesses N universal lossless codes for the N transform coefficients streams.

The rest of the chapter is organized as follows. Section 4.2 reviews and formalizes some results from the ideal coding schemes. Section 4.3 deals with constant stepsize algorithms. Section 4.4 states how both the transformations and the quantization stepsize are adapted, for two different algorithms. Section 4.5 derives the distortion analysis for the two proposed algorithms using adaptive stepsize; the problem of the rates is investigated in section 4.6. Finally, the last section compares the proposed model with numerical results.

4.2 Framework and Background

4.2.1 Quantization Stepsize and Optimal Bit Assignment

Assuming an optimal bit assignment for some transform components y_i, the distortion-rate function for the vectorial signal $\underline{y}$ is at sufficiently high rates

$$\mathrm{E}\,||\underline{\tilde{y}}||_T^2 = \sum_{i=1}^{N} \sigma_{q_i}^2 = Nc2^{-2r}(\prod_{i=1}^{N} \sigma_{y_i}^2)^{1/N} = N\sigma_q^2. \tag{4.1}$$

The corresponding number of bits assigned to the ith component is

$$r_i = r + \frac{1}{2}\log_2 \frac{\sigma_{y_i}^2}{\left(\prod_{i=1}^{N} \sigma_{y_i}^2\right)^{\frac{1}{N}}}. \tag{4.2}$$

Under high resolution assumption, the quantization noise resulting from quantization with stepsize Δ_i is a uniformly distributed random variable (r.v.), with variance $\sigma_{q_i}^2 = \frac{\Delta_i^2}{12}$. A simple way of realizing the optimal bit assignment is thus to quantize all the components with an equal stepsize Δ. If the y_i^q are entropy coded, the bitrate is for Gaussian signals

$$r_i = H(y_i^q) \approx \frac{1}{2}\log_2 2\pi e \sigma_{y_i}^2 - \log_2 \Delta. \tag{4.3}$$

It can then easily be checked that choosing

$$\Delta = \sqrt{2\pi e}2^{-r}(\prod_{i=1}^{N} \sigma_{y_i}^2)^{\frac{1}{2N}} = \sqrt{2\pi e}2^{-r}\det(\text{ diag}\{TRT^T\})^{\frac{1}{2N}} \tag{4.4}$$

yields $\frac{1}{N}\sum_{i=1}^{N} r_i = \frac{1}{N}\sum_{i=1}^{N} H(y_i^q) \approx r$. At high rates, the corresponding distortion-rate function is then[2]

$$D(r) \approx \frac{\Delta^2}{12} \approx \frac{\pi e}{6}2^{-2r}\det(\text{ diag}\{TRT^T\})^{\frac{1}{N}}. \tag{4.5}$$

Relations (4.4) and (4.5) allow therefore to choose a target point (r,D,R) for the transform coding system. Note that this strategy assumes the knowledge of R.

[2] In order to simplify the notations, D, instead of E $||.||^2$, will denote the distortion in the rest of the chapter.

4.2.2 Optimal Transforms

In the unitary case, the optimal transform for Gaussian sources is a KLT V: $VRV^T = \Lambda$, the variances of the transform signals are the eigenvalues λ_i of R.
In the causal case, $\underline{y}_k = L\underline{x}_k = \underline{x}_k - \overline{L}\underline{x}^q_k$, where $\overline{L}\underline{x}^q_k$ is the reference vector. The output $\underline{x}^q_k$ is $\underline{y}^q_k + \overline{L}\underline{x}^q_k$. If we neglect the fact that the prediction uses quantized data, it was shown in chapter 2 that the optimal causal L in terms of coding gain is such that $L\ R\ L^T = \overline{\text{diag}}\,\{\sigma^2_{y_1}, \cdots, \sigma^2_{y_N}\}$, where $\overline{\text{diag}}\,\{\underline{a}\}$ represents a diagonal matrix with diagonal $\underline{a}$. The components y_i are the prediction errors of x_i with respect to the past values of $\underline{x}$, the $\underline{x}_{1:i-1}$, and the optimal coefficients $-L_{i,1:i-1}$ are the optimal prediction coefficients. For both transforms, the high resolution distortion using ECUQ is then

$$D_0(r) \approx \frac{\pi e}{6} 2^{-2r} \left(\det R\right)^{\frac{1}{N}} . \tag{4.6}$$

From (4.4), this distortion corresponds to a quantization stepsize Δ_0 given by

$$\Delta_0 = \sqrt{2\pi e} 2^{-r} (\prod_{i=1}^{N} \sigma^2_{y_i})^{\frac{1}{2N}} = \sqrt{2\pi e} 2^{-r} \left(\det R\right)^{\frac{1}{2N}} \tag{4.7}$$

4.3 Backward Adaptive Algorithms with Fixed Stepsize

From the previous section, assuming that the encoder has the knowledge of the covariance matrix R, it may choose a target rate-distortion point for the system (r,$D(r)$,R). For sufficiently high resolution, this point is determined by choosing, for a given source, a stepsize $\Delta_r = \sqrt{2\pi e} 2^{-2r} \left(\det R\right)^{\frac{1}{2N}}$. Two questions arise regarding the ability of the backward system to converge to this point. Firstly, the decoder should have *a priori* knowledge of the desired Δ_r. Secondly, the estimated transforms should converge to the optimal ones (so that the actual product of the variances $\sigma^2_{y_i}$ is, after convergence, actually $\det R$).
As far as Δ_r is concerned, one shall assume that it is transmitted at the beginning of the coding process as side-information to the decoder. In this sense, the scheme is not fully backward adaptive. The corresponding excess bitrate is small and vanishes in the limits of the data length if the process is stationary; but if the source is time-varying, this may cause a non negligible overhead. The question of the transform convergence will be investigated after we have precisely described the coding algorithm.
Assuming that an estimate

$$\widehat{R}_K = \frac{1}{K} \sum_{i=1}^{K} \underline{x}^q_i \underline{x}^{q\,T}_i \tag{4.8}$$

of the covariance matrix is available at the decoder, the transforms $\widehat{T} = \widehat{V}$, $\widehat{L}$ can be computed so that $\widehat{T}\widehat{R}\widehat{T}^T$ is diagonal. We assume that the first N vectors are sent with very high resolution to the decoder: $\underline{x}^q_i \approx \underline{x}_i,\ i = 1, \cdots, N$. This leads to the following backward adaptive algorithm:

Algorithm $[0]$:

- Initialization: $K = N$.
- Step 1: An estimate of the covariance matrix $\widehat{R}_K = \frac{1}{K}\sum_{i=1}^{K} \underline{x}_i^q \underline{x}_i^{qT}$ is available at both the encoder and the decoder.
- Step 2: A transform $\widehat{T}_K$ is computed such that $\widehat{T}_K \widehat{R}_K \widehat{T}_K^T$ is diagonal, where $\widehat{T}_K$ is either a KLT, either an LDU factorization of $\widehat{R}_K$.
- Step 3: These tranforms are used to transform and quantize the $(K+1)$th vector by $\underline{y}^q_{K+1} = [\widehat{V}_K \underline{x}_{K+1}]_{\Delta_r}$ in the unitary case, or $\underline{y}^q_{K+1} = [\underline{x}_{K+1} - \widehat{\overline{L}}\underline{x}^q_{K+1}]_{\Delta_r}$ in the causal case, where $[.]_{\Delta_r}$ denotes uniform quantization with stepsize Δ_r.
- Step 4: Back to Step 1: the decoder computes then an estimate of the covariance matrix $\widehat{R}_{K+1} = \frac{1}{K+1}(\sum_{i=1}^{N} \underline{x}_i \underline{x}_i^T + \underline{x}^q_{K+1}\underline{x}^{qT}_{K+1})$, from which $\widehat{T}_{K+1}$ can be computed, used to code the $(N+2)$th vector, and so on.

The corresponding block diagram is depicted in figure 4.1.

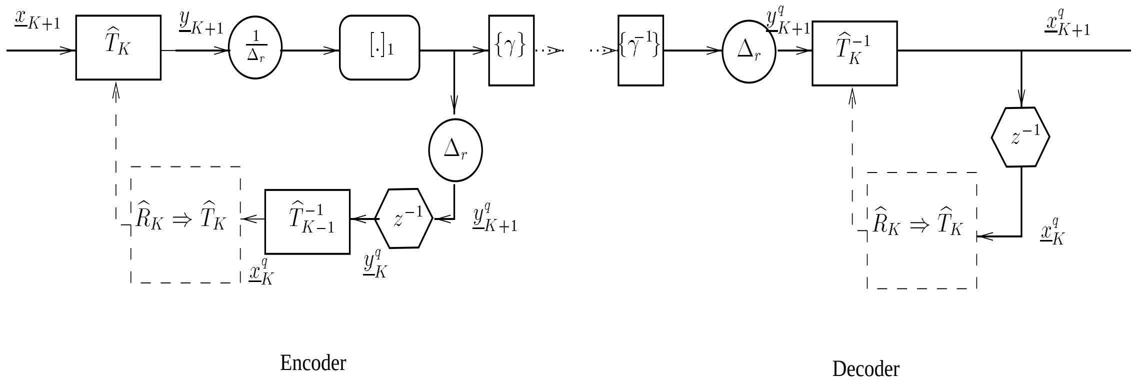


Figure 4.1: Backward adaptive transform coding system with fixed quantization stepsize Δ_r.

For this algorithm, the question is to know whether the transforms will converge or not to the optimal transforms. This algorithm was inspected in the unitary case in [54]; the following conclusions can be drawn.

Assuming on the one hand that the effect of the quantization is to add a zero-mean signal $\underline{z}$ independent of $\underline{x}$ with $\mathrm{E}\,\underline{z}\underline{z}^T = \frac{\Delta^2}{12}I$, the expected covariance matrix $\mathrm{E}\,\underline{x}^q\underline{x}^{qT}$ is $R_{\underline{x}\underline{x}} + \frac{\Delta^2}{12}I$. Since $R_{\underline{x}\underline{x}} + \frac{\Delta^2}{12}I$ and $R_{\underline{x}\underline{x}}$ have the same eigenvectors, the transform converges to the correct transform, resulting in an *universal* system. Universal means here that this performance approaches that of an ideal transform code designed with *a priori* knowledge of the source distribution.

As detailed in section 2.A on the other hand, the difference between $R_{\underline{x}\underline{x}}$ and $R_{\underline{x}^q\underline{x}^q}$ is not precisely a scaled Identity matrix. Moreover, the distribution of $\underline{x}_i^q$ depends on $\widehat{T}_{i-1}$, which in turn depends on the whole sequence $\underline{x}^q_{k=1:i}$. This interdependence renders the analysis of the convergence difficult. Some convergence results are however proven in [54]. Assuming simplifying assumptions such as introducing dithering (whose effect is to make the quantization noise precisely input-independent), or neglecting the stochastic aspect in (4.8), theoretical proofs for the universality of such constrained systems can be established. For the general system described by algorithm $[\mathit{0}]$, it is nevertheless asserted in [54] that convergence should hold as well. One of the aim of this section is to provide empirical evidence that the backward adaptive unitary system actually works.

In the causal case, the same comments can be made regarding the complexity of the convergence analysis. Assuming that $\mathrm{E}\,\underline{x}^q_i\underline{x}^{q\,T}_i$ converges to $R_{\underline{x}\underline{x}} + \frac{\Delta^2}{12}I$, it is necessary that the decoder computes $\widehat{L}_K$ such that $\widehat{L}_K(R_{\underline{x}\underline{x}} + \frac{\Delta^2}{12}I)\widehat{L}_K^T$ is diagonal. From the expression (2.33) of chapter 2, this is precisely the optimal strategy to compute the best causal transform optimized for a closed loop system, at moderate to high rates. Note that equation (2.33) does not assume that the covariance matrix of the quantized data is $R_{\underline{x}^q\underline{x}^q} = R_{\underline{x}\underline{x}} + \frac{\Delta^2}{12}I$; instead, this expression expresses the fact that, when optimized for a closed loop system, the optimal transform can be seen as the optimal prediction matrix for the signal perturbed by a white noise. In the case where the only available estimate is $R_{\underline{x}\underline{x}} + \frac{\Delta^2}{12}I$, computing the corresponding LDU factorization of $\widehat{R}_{\underline{x}^q\underline{x}^q}$ is equivalent to finding the correct prediction matrix. Thus, at least at moderate to high rates, the causal system should be universal as well.

Figure 4.2 plots the actual rate-distortion functions obtained with algorithm $[\mathit{0}]$, for the same source as in 2.6.3 (decreasing variances). Sequences of 10^4 vectors were backward adaptively transform coded as described in algorithm $[\mathit{0}]$, for several stepsizes Δ_r. For each stepsize, the resulting distortion and entropy were measured for the whole sequence; the experiment was repeated 10 times. Comparing with figure 2.17, which plots the actual rate-distortion functions for transforms designed with *a priori* knowledge of the statistics of the source, the similarity is apparent. In particular, the system converges even at low rates, when the quantization noise is large.

An interesting question for this algorithm would be the following. Let us assume that the decoder stops adapting the transforms after a certain amount of K vectors. Then what would be the rate required to code (asymptotically in the data length) the resulting source ? This question will not be addressed here, but an analysis provided in chapter 6 deals with the same interrogation in the framework of lossless coding.

The algorithm $[\mathit{0}]$ as described above suffers from a drawback when the source is not stationary but locally stationary. Quantizing with the same Δ_r may cause unacceptable changes in rate if the variances of the x_i vary w.r.t. the quantization stepsize; the distortion is fixed, but the SNR will vary as well.

To precise this, let us assume a piecewise stationary vector source, whose covariance matrix R_1 changes to R_2 after a certain time. The encoder may have the knowledge of R_1; according to some rate-distortion trade-off objective (r_1, D_1), it may choose consequently for the source a convenient stepsize $\Delta_{r_1} = \sqrt{2\pi e}2^{-r_1}(\det R_1)^{\frac{1}{2N}}$. This stepsize should be transmitted to the decoder at the beginning of the backward adaptive coding process. Assuming the stationarity period long enough for the process to converge,

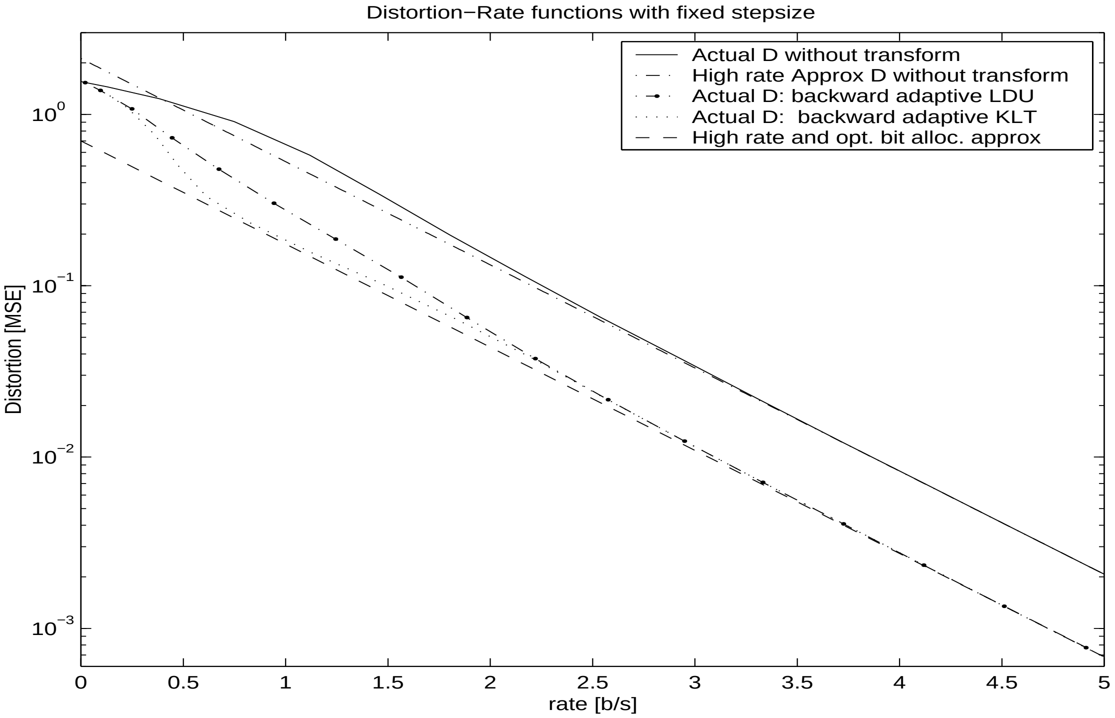


Figure 4.2: Distortion-rate functions for the KLT and the LDU using a backward adaptive algorithm with constant quantization stepsize. N=3 and $\rho = 0.9$.

the system will then work at a point ($D_1 = \frac{\Delta_1^2}{12}$,$r_1$,$R_1$) of the rate-distortion function of R_1. If now the statistics of the source change (covariance matrix R_2), the system working with stepsize Δ_{r_1} will yield after convergence the same distortion D_1; the rate actually required to entropy code the source will become

$$r_2 = \frac{1}{2}\log_2 2\pi e \frac{(\det R_2)^{\frac{1}{N}}}{\Delta_{r_1}^2} = r_1 + \frac{1}{2}\log_2 \left(\frac{\det R_2}{\det R_1}\right)^{\frac{1}{N}} . \tag{4.9}$$

Thus, the rate will change accordingly to the determinant of the covariance matrix of the source. Moreover, the distortion D_1 may be acceptable for R_1, but not for R_2 : consider a single scalar source x with variance σ_x^2. The operational rate-distortion function of this uniformly quantized and entropy coded signal is $d = \frac{\Delta^2}{12} = \frac{\pi e}{6} 2^{-2r}\sigma_x^2$. If the variance of the source is time-varying, it is then more relevant to guarantee the SNR $\frac{\sigma_x^2}{d}$ to be constant rather than d only. Thus, it seems interesting to find a solution which keeps the relation of the distortion to R constant (or equivalently, which keeps the asymptotic rate constant); such a coding scheme should therefore converge to the point $D_2 = \frac{\pi e}{6} 2^{-2r_1}(\det R_2)^{\frac{1}{N}}$.

Possible solutions to that problem exist for algorithms using a fixed stepsize. A more convenient stepsize may for example be retransmitted to the decoder accordingly to the source variations, but this results in some overhead. This side information may be avoided if, after convergence of the process, both the encoder and the decoder change in unison the stepsize according to the new estimate of R_2; this solution lengthen however the time by which the desired performance are achieved. One may therefore try to design algorithms which converge directly to the desired rate-distortion function point of the source, by keeping the target rate constant. This is the aim of the next section, where both the quantization stepsize and the transforms are backward adaptive.

4.4 Backward Adaptive Algorithms with Adaptive Stepsize

4.4.1 Framework

For a nonstationary input, the variances of the sources are time variable; the problem of (backward) adapting the quantization stepsize is very similar if the input is stationary, or locally stationary, with unknown variance. The operation of an adaptive scalar quantizer is therefore of the general form $\Delta_i = \phi \widehat{\sigma}^2_K$, where $\widehat{\sigma}^2_K$ is the variance estimate at time instant K, and ϕ is some constant [14]. We shall thus assume a stationary vector source with unkown covariance matrix R. This source is tranform coded in a backward adaptive manner, and neither the encoder nor the decoder has *a prori* knowledge of R. The transforms and the quantization stepsize will be adapted periodically in unison at the encoder and at the decoder; no side information is therefore required to transmit any coding parameters. The only *a priori* information shared by the encoder and the decoder is the target rate r_0 at which the system should work. Assuming sufficiently high resolution, the goal is then for the system to converge to the rate-distortion point (r_0,D_0,R): $D_0(r_0) = \frac{\pi e}{6} 2^{-2r_0} (\det R)^{\frac{1}{N}}$, which is for this transform coded source the best achievable rate-distortion point at rate r_0.

We now propose two algorithms updating T and Δ by means of the data available at the decoder only. In addition to the assumptions expressed in the Introduction, the first N vectors are assumed to be quantized with very high resolution and sent (without being transformed) to the decoder.

Algorithm $[\mathit{1}]$:

- Initialization: $K = N$.
- Step 1: An estimate of the covariance matrix $\widehat{R}_K = \frac{1}{K} \sum_{i=1}^{K} \underline{x}^q_i \underline{x}^{qT}_i$ is available at both the encoder and the decoder.
- Step 2: A transform $\widehat{T}_K$ is computed such that $\widehat{T}_K \widehat{R}_K \widehat{T}^T_K$ is diagonal, where $\widehat{T}_K$ is either a KLT, either a LDU factorization of $\widehat{R}_K$, and a stepsize $\widehat{\Delta}^{[\mathit{1}]}_K$ is computed by

$$\widehat{\Delta}^{[\mathit{1}]}_K = \sqrt{2\pi e} 2^{-r_0} \det(\widehat{T}_K \widehat{R}_K \widehat{T}^T_K)^{\frac{1}{2N}}. \tag{4.10}$$

- Step 3: These parameters are used to transform and quantize the $(K+1)$th vector by $\underline{y}^q_{K+1} = [\widehat{V}_K \underline{x}_{N+1}]_{\widehat{\Delta}^{[\mathit{1}]}_K}$ in the unitary case, or $\underline{y}^q_{K+1} = [\underline{x}_{K+1} - \widehat{\overline{L}} \underline{x}^q_{K+1}]_{\widehat{\Delta}^{[\mathit{1}]}_K}$ in the causal case, where $[.]_\Delta$ denotes uniform quantization with stepsize Δ. The expected distortion for the $(K+1)$th vector is then $D^{[\mathit{1}]}(K+1) = \mathrm{E}\, \widehat{\Delta}^{[\mathit{1}]^2}_K / 12$.
- Step 4: Back to Step 1: the decoder computes then an estimate of the covariance matrix $\widehat{R}_{K+1} = \frac{1}{K+1} (\sum_{i=1}^{N} \underline{x}_i \underline{x}^T_i + \underline{x}^q_{K+1} \underline{x}^{qT}_{K+1})$, from which $\widehat{T}_{K+1}$ and $\widehat{\Delta}_{K+1}$ can be computed, used to code the $(N+2)$th

vector, and so on.

Algorithm $[2]$:

A simple improvement to the previous algorithm can be made by using the results regarding uniform quantization of Gaussian sources evocated above. For Gaussian vectors $\underline{y}_K$, quantized with the same (constant) stepsize Δ, it can be shown (see 2.A) that

$$\mathrm{E}\,\underline{y}_i^q \underline{y}_i^{qT} = R_{\underline{y}^q \underline{y}^q} = R + \frac{\Delta^2}{12} I + B, \text{ where } B \to 0 \text{ elementwise as } \Delta \to 0. \tag{4.11}$$

In the previous algorithm now, if the stepsize converges to some stepsize $\Delta_\infty(T)$, one may expect that the estimate of the covariance matrix converges to some matrix close to $R + \frac{\Delta_\infty^2(T)}{12} I$. The numerical evaluations of the previous section indicate convergence even for large quantization stepsizes. A better estimate of R may therefore be computed after a certain amount of vectors, say N_1, by substracting $\frac{\widehat{\Delta}_K^2}{12} I$ to the current estimate of R. This correction on the estimate of the second order moment of the data by their quantized version is usually referred to as Sheppard's correction [72]. Except from this difference concerning $\widehat{R}$, the steps of algorithm $[2]$ are the same as in algorithm $[1]$.

The corresponding block diagram of these two algorithms is depicted in figure 4.3.

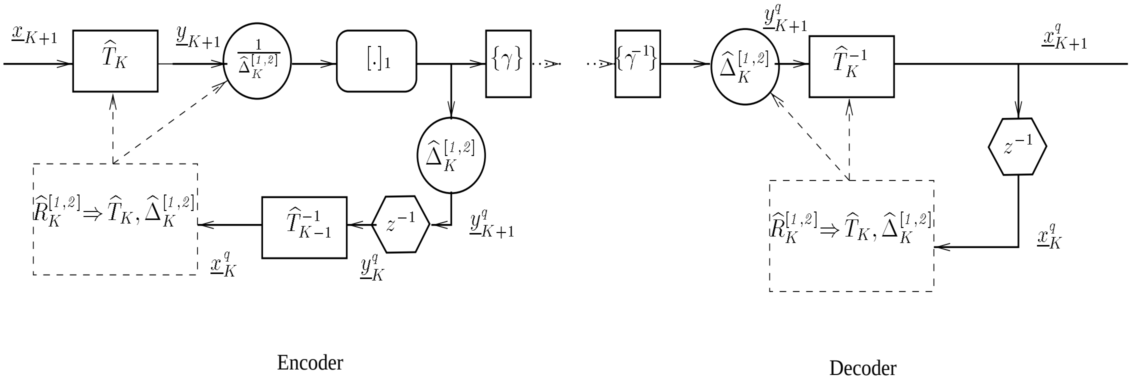


Figure 4.3: Backward adaptive transform coding system with adaptive quantization stepsize.

4.4.2 Proposed Analysis

The convergence analysis for these algorithms will seek to determine if the corresponding distortions converge or not to the target rate r_0 and distortion D_0. Moreover, we will try to model the behaviour of the distortion *vs* K. The proposed analysis will retain only first order perturbation; therefore, it does not claim to establish rigorous convergence proofs.

Two preliminary steps required to analyze the distortion are detailed in the rest of this section. First, we should precisely evaluate the respective contribution of the estimation and quantization noises for the estimates of the covariance matrix. This lead to a handful of perturbation terms. Second, we precise the relationship between the distortion of interest and these perturbation terms.

The estimate of the covariance matrix for the second algorithm can be expressed as

$$\begin{aligned}
\widehat{R}_K^{[2]} &= \tfrac{1}{K}\Big(\sum_{i=1}^{N}\underline{x}_i\underline{x}_i^T + \sum_{i=N+1}^{N_1}\underline{x}_i^{[1]q}\underline{x}_i^{[1]q^T} + \underline{\sum_{i=N_1+1}^{K}\underline{x}_i^{[2]q}\underline{x}_i^{[2]q^T}}\Big) - \underline{\frac{\widehat{\Delta}_{K-1}^{[2]^2}}{12}I} \\
&= \tfrac{1}{K}\left(NR + \sum_{i=N+1}^{N_1}(R + D^{[1]}(i)I) + \underline{\sum_{i=N_1+1}^{K}(R + D^{[2]}(i)I)}\right) \\
&\quad + \tfrac{1}{K}\left(\sum_{i=1}^{N}\Delta R_i^{(1)} + \sum_{i=N+1}^{N_1}\Delta R_i^{q[1](1)} + \underline{\sum_{i=N_1+1}^{K}\Delta R_i^{q[2](1)}}\right) - \underline{\frac{\widehat{\Delta}_{K-1}^{[2]^2}}{12}I}
\end{aligned} \tag{4.12}$$

where $D^{[1,2]}(i)$ denotes the distortion obtained for the ith vector, and where we used the following notation:

- superscript $^{[j]}$ refers to algorithm $[j]$,

- superscript ^{q} refers to quantization,

- superscript $^{(1)}$ refers to estimation noise occuring by estimating a covariance matrix R by the estimate $\underline{x}_i\underline{x}_i^T = \widehat{R}^{(1)} = R + \Delta R^{(1)}$,

- subscript $_K$ refers to the total number of vectors available at the decoder (except indeed from $\underline{x}_i$, which denotes the ith vector).

The corresponding estimate for the first algorithm $\widehat{R}_K^{[1]}$ can also be computed from (4.12), where in this case the underlined terms vanish.

By writing $\frac{\widehat{\Delta}_{K-1}^{[2]^2}}{12} = D^{[2]}(K) + \delta D^{[2]}(K)$, the estimate (4.12) can also be written as $\widehat{R}_K^{[2]} = R + \Delta R_K^{[2]}$, with

$$\begin{aligned}
\Delta R_K^{[2]} &= \underbrace{\left[\frac{1}{K}\left(\sum_{i=N+1}^{N_1} D^{[1]}(i) + \sum_{i=N_1+1}^{K} D^{[2]}(i)\right) - D^{[2]}(K)\right] I}_{\Delta R_{K,det}^{[2]}} \\
&\quad \underbrace{+\frac{1}{K}\left(\sum_{i=1}^{N}\Delta R_i^{(1)} + \sum_{i=N+1}^{N_1}\Delta R_i^{q[1](1)} + \sum_{i=N_1+1}^{K}\Delta R_i^{q[2](1)}\right) - \delta D^{[2]}(K),}_{\Delta R_{K,sto}}
\end{aligned} \tag{4.13}$$

where $\Delta R_{K,det}^{[2]}$ is a deterministic diagonal matrix, and $\Delta R_{K,sto}^{[2]}$ is a stochastic matrix. The update of the transform (to simplify the notations, the subscript $^{[2]}$ will be omitted for $\widehat{T}_K^{[2]}$) is then computed so that $\widehat{T}_K\widehat{R}_K^{[2]}\widehat{T}_K^T$ is diagonal, and the updated stepsize $\widehat{\Delta}_K^{[2]} = \sqrt{2\pi e}2^{-r_0}\det\left(\widehat{T}_K\widehat{R}_K^{[2]}\widehat{T}_K^T\right)^{\frac{1}{2N}}$ is used to quantize the $(K+1)$th transform vector. For sufficiently small stepsizes, the expected distortion is then

$$D^{[2]}(K+1) \approx \mathrm{E}\,\frac{\widehat{\Delta}_K^{[2]^2}}{12} \approx \frac{\pi e}{6}2^{-2r_0}\det(\widehat{T}_K\widehat{R}_K^{[2]}\widehat{T}_K^T)^{\frac{1}{N}}, \tag{4.14}$$

The corresponding distortion $D^{[1]}(K+1)$ for algorithm $[1]$ can be computed by simplifying in $D^{[2]}(K+1)$ the vanishing terms of $\widehat{R}^{[2]}$, see (4.12).

Using the unimodularity property of the transforms and considering $\Delta R_K^{[2]}$ in (4.13) as a perturbation term uppon R, one should compute in both unitary and causal cases (tr denotes the trace operator, see the derivation in appendix (4.A))

$$\begin{aligned} D^{[2]}(K+1) &\approx \frac{\pi e}{6} 2^{-2r_0} (\det \widehat{R}_K^{[2]})^{\frac{1}{N}} \\ &\approx D_0[1+\tfrac{1}{N}\,\mathrm{E}\ \mathrm{tr}\{\Delta R_K^{[2]} R^{-1}\}+\tfrac{1}{2N^2}\,\mathrm{E}\,(\ \mathrm{tr}\{\Delta R_K^{[2]} R^{-1}\})^2-\tfrac{1}{2N}\,\mathrm{E}\ \mathrm{tr}\{\Delta R_K^{[2]} R^{-1} \Delta R_K^{[2]} R^{-1}], \end{aligned} \tag{4.15}$$

The corresponding distortion for algorithm $[1]$ can be computed from $D^{[2]}$ by inspecting the vanishing terms in $\Delta R_K^{[2]}$ through (4.12).

4.5 Distortion Analysis

In order to compute the three expectations in (4.15), we can describe the r.v.s involved in (4.13) as follows. The elementary terms $\{\Delta R_i^{(1)}\}$ corresponds to ”one-shot” estimates of R based on a single observation. Since the vectors $\underline{x}_K$ are i.i.d., so is $\Delta R_i^{(1)}$. The elementary terms $\{\Delta R_i^{q[1,2](1)}\}$ correspond to ”one-shot” estimates of $R+\ \mathrm{E}\,(\widehat{\Delta}_{i-1}^{[1,2]^2}/12)I$ which, from (4.11), can be approximated as $R+\ D^{[1,2]}(i)I$. These terms are indeed not identically distributed. They are neither independent since $\Delta R_i^{q[1,2](1)}$ depends on $\widehat{\Delta}_{i-1}^{[1,2]}$, which depends on $\widehat{R}_{i-1}^{[1,2]}$, which in turn depends on $\Delta R_i^{q[1,2](1)}$. However, we assume that this is the case, since this dependence concerns only the noise part of the quantized vectors. Because of the quantization noise, the vectors $\underline{x}_K^q$ are not strictly Gaussian; for sufficiently high resolution, we assume that this is however the case.

The following result (see appendix (3.A)) is now necessary to establish (4.15)[3]. Let $\Delta R_l^{(1)} = R_l = \underline{x}_i \underline{x}_i^T$ be the (symmetric) estimate of some $R_l = [\underline{r}_{l_1} \ldots \underline{r}_{l_N}]$ by means of one real zero mean Gaussian vector $\underline{x}_i$, with $\mathrm{E}\,\underline{x}_i \underline{x}_i^T = R_l$. Then it can be shown that $\Delta R_l^{(1)}$ is a zero mean r.v., and that among the N^2 blocks of $\mathrm{E}\,\mathrm{vec}\Delta R_l^{(1)} \mathrm{vec}^T \Delta R_l^{(1)}$, the (i,j)th block

$$(\,\mathrm{E}\,\mathrm{vec}\Delta R_l^{(1)} \mathrm{vec}^T \Delta R_l^{(1)})_{block_{(i,j)}} = (R_l \otimes R_l)_{block_{(i,j)}} + \underline{r}_{l_j} \underline{r}_{l_i}^T, \tag{4.16}$$

where $\otimes$ denotes the Kronecker product. If now $R_l = R + D_l I$, with I denotes Identity and D_l a scalar the previous expression may, for correlated sources, be approximated as

$$\mathrm{E}\,\mathrm{vec}\Delta R_l^{(1)} \mathrm{vec}^T \Delta R_l^{(1)} \approx 2R_l \otimes R_l \approx 2\,[R \otimes R + D_l (R \otimes I + I \otimes R)]\,. \tag{4.17}$$

The first term of (4.15) may be written as

$$\begin{aligned} \tfrac{1}{N}\,\mathrm{E}\ \mathrm{tr}\{\Delta R_K^{[2]} R^{-1}\} &= \tfrac{1}{N} \left(\mathrm{tr}\{\Delta R_{K,det}^{[2]} R^{-1}\} + \underbrace{\mathrm{E}\ \mathrm{tr}\{\Delta R_{K,sto}^{[2]} R^{-1}\}}_{0} \right) \\ &\approx \tfrac{1}{N} \left[\tfrac{1}{K}\Sigma_{tot} - D^{[2]}(K)\right]\ \mathrm{tr}\{R^{-1}\}, \end{aligned} \tag{4.18}$$

[3] The derivations involved in the computations of (4.15) are only outlined in this section; the details are reported in sec. 4.B

with $\Sigma_{tot} = \sum_{i=N+1}^{N_1} D^{[1]}(i) + \sum_{i=N_1+1}^{K} D^{[2]}(i)$. The second term leads to

$$\begin{aligned}\frac{1}{2N^2}\,\mathrm{E}\,(\,\mathrm{tr}\,\{\Delta R_K^{[2]} R^{-1}\})^2 &= \frac{1}{2N^2}\,\mathrm{E}\,(\,\mathrm{tr}\,\{R^{-\frac{1}{2}}\Delta R_K^{[2]} R^{-\frac{1}{2}}\})^2 \\ &\approx \frac{1}{2N^2}(\mathrm{vec}^T R^{-\frac{1}{2}})(R^{-\frac{1}{2}} \otimes I) \quad \underbrace{\mathrm{E}\,\mathrm{vec}\Delta R_K^{[2]}\mathrm{vec}^T \Delta R_K^{[2]}}_{\mathrm{vec}\Delta R_{K,det}^{[2]}\mathrm{vec}^T \Delta R_{K,det}^{[2]} +!\ \mathrm{E}\,\mathrm{vec}\Delta R_{K,sto}^{[2]}\mathrm{vec}^T \Delta R_{K,sto}^{[2]}} \quad (R^{-\frac{1}{2}} \otimes I)\mathrm{vec} R^{-\frac{1}{2}}\end{aligned} \tag{4.19}$$

where the term corresponding to the deterministic part can be computed using the fact that $\Delta R_{K,det}^{[2]}$ is diagonal. The stochastic term in (4.19) generates, according to (4.13), four terms, which can be computed using (4.17). The second term in (4.15) leads finally to

$$\frac{1}{2N^2}\,\mathrm{E}\,(\,\mathrm{tr}\,\{\Delta R_K^{[2]} R^{-1}\})^2 \approx \frac{1}{KN} + \,\mathrm{tr}\,\{R^{-1}\}(\frac{2\Sigma_{tot}}{K^2N^2}) + \frac{\left(\mathrm{tr}\,\{R^{-1}\}\right)^2}{2N^2}\left[\left(\frac{\Sigma_{tot}}{K} - D^{[2]}(K)\right)^2\right], \tag{4.20}$$

where for the purpose of this first order analysis, only the dominating terms have been retained. Concerning the third term of (4.15), let G be $R^{-\frac{1}{2}}\Delta R_K^{[2]} R^{-\frac{1}{2}}$. Then we have vec$G = (R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})vec\Delta R_K^{[2]}$, and we get

$$\begin{aligned}-\frac{1}{2N} E\,\mathrm{tr}\,\{\Delta R_K^{[2]} R^{-1}\Delta R_K^{[2]} R^{-1} &= \mathrm{E}\, - \frac{1}{2N}\,\mathrm{tr}\,\{GG\} \\ &= -\frac{1}{2N}\,\mathrm{E}\,\,\mathrm{tr}\,\{\mathrm{vec}G\mathrm{vec}^T G\} \\ &= -\frac{1}{2N}\,\mathrm{E}\,\,\mathrm{tr}\,\{(R^{-\frac{1}{2}}\otimes R^{-\frac{1}{2}})\,\mathrm{E}\,\mathrm{vec}\Delta R_K^{[2]}\mathrm{vec}^T \Delta R_K^{[2]}(R^{-\frac{1}{2}}\otimes R^{-\frac{1}{2}})\}\end{aligned} \tag{4.21}$$

where again, the arising terms can be computed using (4.17).

Finally, the distortion occuring with the second algorithm can be approximated by the recursive expression

$$D^{[2]}(K+1) \approx D_0\left[1 + \frac{1}{K}(\frac{1}{N} - N) + \,\mathrm{tr}\,\{R^{-1}\}\left(\frac{1}{N}\left[\frac{1}{K}\left(\sum_{i=N+1}^{N_1} D^{[1]}(i) + \sum_{i=N_1+1}^{K} D^{[2]}(i)\right) - D^{[2]}(K)\right]\right)\right]. \tag{4.22}$$

Inspecting the vanishing terms in (4.12), we obtain then the following recursive expression for the algorithm without correction

$$D^{[1]}(K+1) \approx D_0\left[1 + \frac{1}{K}(\frac{1}{N} - N) + \frac{\mathrm{tr}\,\{R^{-1}\}}{KN}\left(\sum_{i=N+1}^{K} D^{[1]}(i)\right)\right]. \tag{4.23}$$

On the one hand, the recursive expression (4.22) shows that the algorithm based on the Sheppard's correction should, as $K \to \infty$, converge to the target distortion D_0,

$$D_\infty^{[2]} \approx \mathrm{E}\,\frac{\Delta_\infty^{[2]^2}}{12} \approx D_0, \tag{4.24}$$

The corresponding stesize should converge to

$$\mathrm{E}\,\Delta_\infty^{[2]} \approx \Delta_0 \approx \sqrt{2\pi e}(\det R)^{\frac{1}{2N}}. \tag{4.25}$$

On the other hand, the model provided by (4.23) does not converge to D_0 but to some $D_\infty^{[1]} > D_0$, which can easily be computed as

$$D_\infty^{[1]} \approx \mathrm{E}\,\frac{\Delta_\infty^{[1]^2}}{12} \approx \frac{D_0}{1 - D_0 \frac{\mathrm{tr}\{R^{-1}\}}{N}}. \tag{4.26}$$

Accordingly, the quantization stepsize should converge to

$$\Delta_\infty^{[1]} \approx (12 D_\infty^{[1]})^{\frac{1}{2}} \approx 2 \left(\frac{\sqrt{3} D_0}{1 - D_0 \frac{\mathrm{tr}\{R^{-1}\}}{N}} \right)^{\frac{1}{2}}. \tag{4.27}$$

Note that at low rates, the convergence of $\Delta_\infty^{[2]}$ to Δ_0 does not guarantee $D_\infty^{[2]}$ to be axactly D_0 because the quantizer's rate-distortion performance factor deviate from $\frac{\pi e}{6}$.

4.6 Rate Analysis

This section analyzes the bitrate required to entropy code the transform signals as $K \to \infty$. These results are therefore asymptotic in the data length.

4.6.1 Algorithm with Sheppard's correction

For the algorithm using the correction on the second order moment estimate, the rate is

$$\begin{aligned} r_{(T)}^{[2]} &= \tfrac{1}{N} \sum_{i=1}^{N} H(y_i^q) \\ &\approx \tfrac{1}{2N} \sum_{i=1}^{N} \log_2 2\pi e \sigma_{y_i,\infty}^2 - \log_2 \Delta_0 \\ &\approx r_0 + \tfrac{1}{2N} \log_2 \frac{\prod_{i=1}^{N} \sigma_{y_i,\infty}^2}{\det R}, \end{aligned} \tag{4.28}$$

where $\sigma_{y_i,\infty}^2$ are the variances of the transform signals obtained by using the transform based on the asymptotic estimate $\widehat{R}_\infty^{[2]}$, which in this case is R. Thus, the estimated KLT and LDU should converge to the optimal transforms. The variances of the transform signals in the unitary case are then λ_i and

$$r_{(V)}^{[2]} = r_0. \tag{4.29}$$

In the causal case, one should account for the fact that the reference signal is computed by means quantized data. The actual prediction error variances $\sigma_{y_i}^{'2}$ are greater than the optimal ones $\sigma_{y_i}^2$ due to a quantization noise feedback similar to that occuring in DPCM, and from (2.69), are approximately given by

$$\prod_{i=1}^{N} \sigma'^2_{y_i} \approx \det(R) \left(1 + D_0(r) \sum_{i=1}^{N} \left(\frac{1}{\lambda_i} - \frac{1}{\sigma_{y_i}^2} \right) \right). \tag{4.30}$$

This traduces, for a given distortion, by an increase in rate approximately given by

$$r^{[2]}_{(L)} \approx r_0 + \frac{D_0}{2N \ln 2} \sum_{i=1}^{N} \left(\frac{1}{\lambda_i} - \frac{1}{\sigma_i^2} \right). \tag{4.31}$$

As a conclusion, though the target distortion is reached in both cases, the unitary approach yields to lowest asymptotic rate because of the noise feedback occuring in the causal approach. From the analysis of chapter 2 section 2.5.2 however, these effects are noticeable at low rates only; for moderate to high target rates

$$r^{[2]}_{(L)} \approx r^{[2]}_{(V)} = r_0. \tag{4.32}$$

4.6.2 Algorithm without Sheppard's correction

For the algorithm $[1]$ now, one should compute

$$r^{[1]}_{(T)} = \frac{1}{N} \sum_{i=1}^{N} H(y_i^q) \approx \frac{1}{2N} \sum_{i=1}^{N} \log_2 2\pi e \sigma^2_{y_i,\infty} - \log_2 \Delta_\infty \tag{4.33}$$

where, this time, the $\sigma^2_{y_i,\infty}$ are the variances of the transform signals obtained by using the transform based on the asymptotic estimate $\widehat{R}^{[1]}_\infty \approx R + \frac{\Delta^2_\infty}{12} I$. In the unitary case, since a KLT of R is also a KLT of $R + \frac{\Delta^2_\infty}{12} I$, the $\sigma^2_{y_i,\infty}$ should again be equal to the λ_i. Using (4.26), we obtain

$$r^{[1]}_{(V)} \approx r_0 - \frac{D_0}{2N \ln 2} \operatorname{tr}\{R^{-1}\}. \tag{4.34}$$

In the causal case, the noise feedback in (4.30) involves this time $\Delta_\infty = 12 D_\infty^{1/2}$, and computing (4.33) yields

$$r^{[1]}_{(L)} \approx r^{[2]}_{(L)} - \frac{D_0}{2N \ln 2} \operatorname{tr}\{R^{-1}\}. \tag{4.35}$$

Thus, the effect of not using the Sheppard correction in the backward adaptive algorithms is, for both transforms, to deplace the actual rate-distortion point ($r_0 - \delta r_0$,$D_0 + \delta D_0$,R) from the target point (r_0,D_0,R) by a rate

$$\delta r_0 \approx \frac{D_0}{2N \ln 2} \operatorname{tr}\{R^{-1}\} \tag{4.36}$$

and from (4.26), by a distortion

$$\delta D_0 \approx D_0^2 \operatorname{tr}\{R^{-1}\}/N. \tag{4.37}$$

From (4.36) and (4.37), these mismatches vanish in the limit of high resolution systems (small target distortion/high target rates). Thus, the effects of the Sheppard's correction become undetectable in the limit of high rates, in which case the behaviour of the algorithms $[1]$ and $[2]$ become equivalent.

4.7 Numerical Results

For the simulations, the data are real Gaussian i.i.d. vectors with covariance matrix $R = HR_{AR1}H^T$. R_{AR1} is the covariance matrix of an AR(1) process with $\rho = 0.9$. H is a diagonal matrix whose ith entry is $(N-i+1)^{1/3}$, $N = 3$. The target rate is 3 b/s. Algorithms $[\mathit{1}]$ and $[\mathit{2}]$ were then implemented as detailed in section 4.4. For each estimate $\widehat{R}^{[\mathit{1},\mathit{2}]}$, the corresponding stepsize $\widehat{\Delta}^{[\mathit{1},\mathit{2}]}$ was computed as in (4.10), and the distortion was estimated by $\frac{\widehat{\Delta}^{[\mathit{1},\mathit{2}]^2}}{12}$. This experiment was repeated 200 times.

- Figure 4.4 plots the averaged observed distortions for the KLT and the LDU versus K for algorith $[\mathit{1}]$ (without Sheppard's correction). The theoretical model is given by (4.23), and the theoretic asymptotic distortion by (4.26). As commented in the text, this distortion should be the same for both transforms, because it close to $\mathrm{E}\,\frac{\widehat{\Delta}^{[\mathit{1}]^2}}{12}$, and this adaptive stepsize, as computed by (4.10), is the same for the KLT and the LDU, because they are unimodular. The target distortion is given by (4.6).
  It can be observed that the estimated distortion converges to theoretical limit (4.26). The excess in distortion is due to the convergence of the stepsize to Δ_∞, as given by (4.27), instead of to $(12D_0)^{\frac{1}{2}}$.

- Similar results are shown in figure 4.5 for the algorithm $[\mathit{2}]$, where the Sheppard's correction is applied after $N_1 = 60$ vectors. The discontinuity is caused by the substraction of $\widehat{\Delta}_{N_1}^{[\mathit{1}]^2}/12\ I$ from the estimate $\widehat{R}_{N_1}^{[\mathit{1}]}$; this decreases the determinant of $\widehat{R}_{N_1}^{[\mathit{1}]}$, and $\widehat{\Delta}_{N_1+1}^{[\mathit{2}]^2}$ is consequently smaller than $\widehat{\Delta}_{N_1}^{[\mathit{1}]^2}$. The theoretical model for $D^{[\mathit{2}]}\ vs\ K$ is given by (4.22). Discontinuity appears clearly after N_1 vectors.

- Figures 4.6 and 4.7 plot the results for the two algorithms for a target rate of $r_0 = 4$ b/s. It can be observed that the behaviour of the two algorithms is similar because the resolution is sufficiently high (the mismatch δD_0 becomes negligible). At higher rates, the stepsizes converge to Δ_0 and the distortions to D_0 for both algorithms. Comparing fig 4.7 and 4.5, the discontinuity due to the Sheppard's correction is decreased, because this correction vanishes in the limit of small distortions.

- Finally, the convergence of the two algorithms at a lower rate (2.3 b/s) is presented in figure 4.8 and 4.9 respectively. It can be observed that $\frac{\widehat{\Delta}^{[\mathit{1},\mathit{2}]^2}}{12}$ does not converge exactly to the theoretical bounds. A this rate and beyond, the high resolution approximations assumed in the theoretical analyses become less accurate. The largest mismatch for both algorithms occurs for the LDU; for this transform, the noise feedback makes the quantization noise and the input the most correlated at low rates, so that the perturbation deviates from a scaled diagonal identity. Moreover, the actual distortion may be different from $\frac{\widehat{\Delta}^{[\mathit{1},\mathit{2}]^2}}{12}$ for both transforms.

Summarizing these results, the proposed analysis of the convergence behaviour of $\mathrm{E}\,\frac{\widehat{\Delta}_K^{[\mathit{1},\mathit{2}]^2}}{12}$ match adequatly the actual convergence behaviour for rates higher than approximately 2.5 bits per sample.

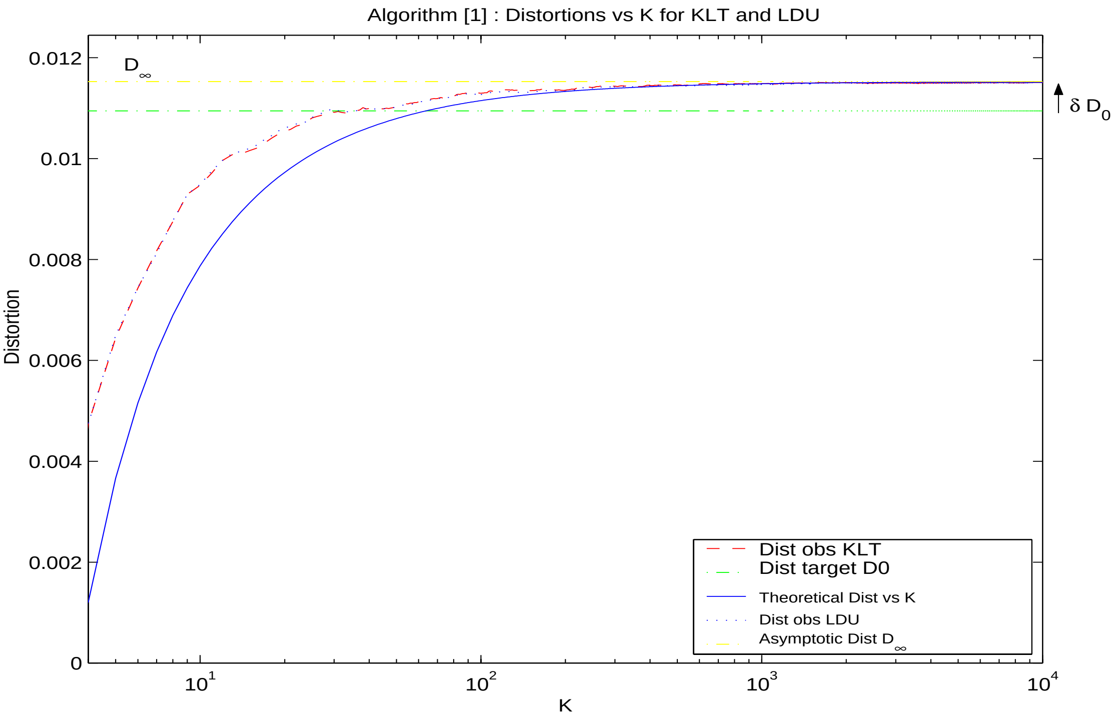


Figure 4.4: Distortions for algorithm $[1]$ *vs* K, $r_0 = 3$ b/s.

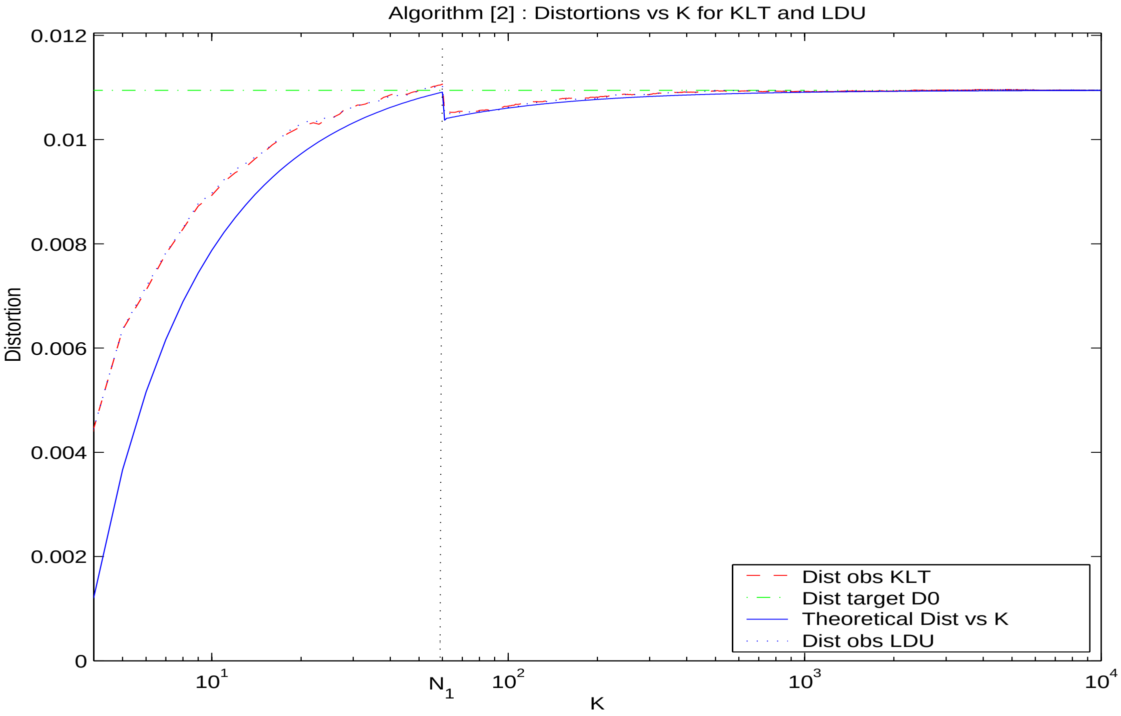


Figure 4.5: Distortions for algorithm $[2]$ *vs* K, $r_0 = 3$ b/s.

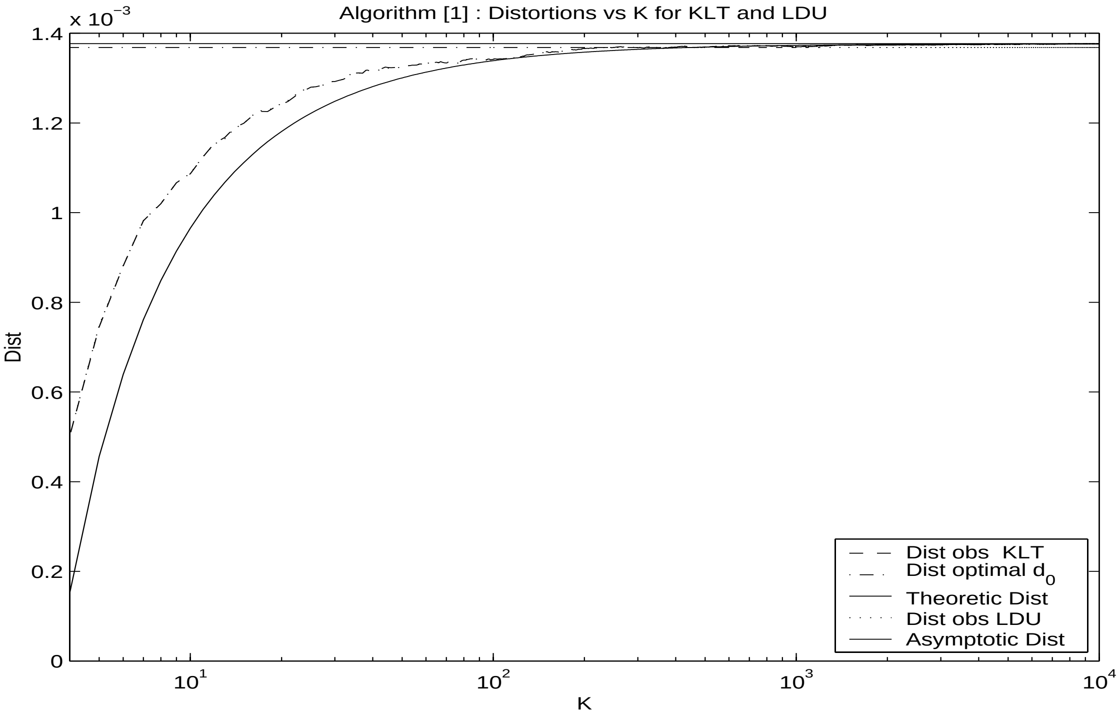


Figure 4.6: Distortions for algorithm $[1]$ *vs* K, $r_0 = 4$ b/s.

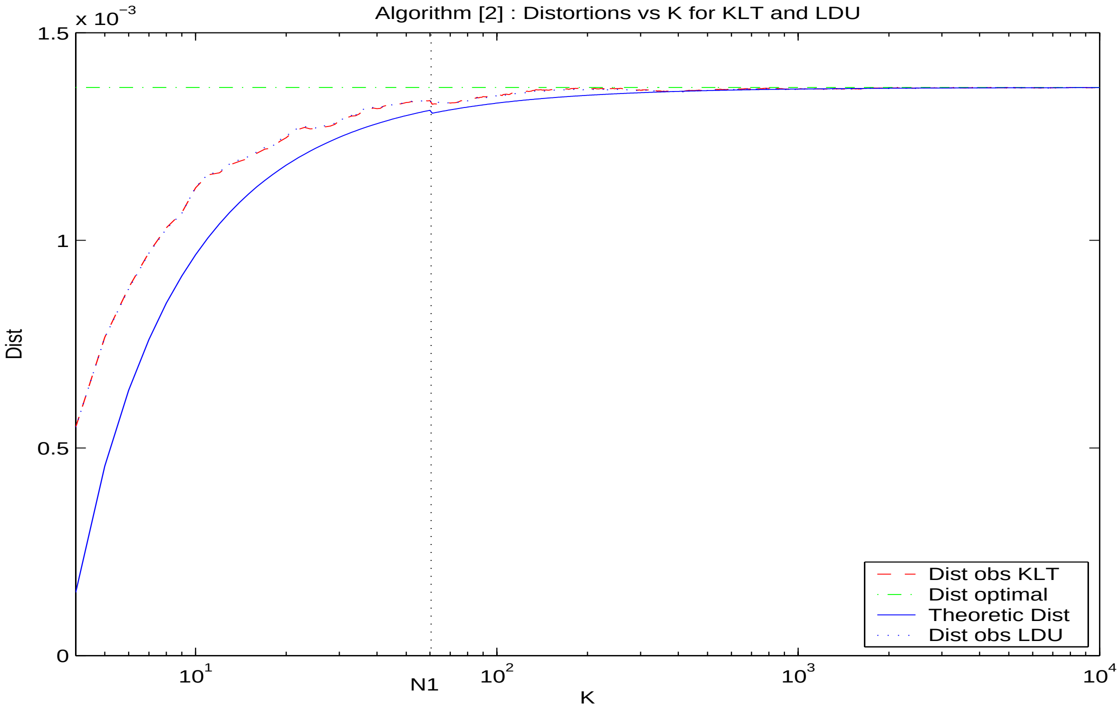


Figure 4.7: Distortions for algorithm $[2]$ *vs* K, $r_0 = 4$ b/s.

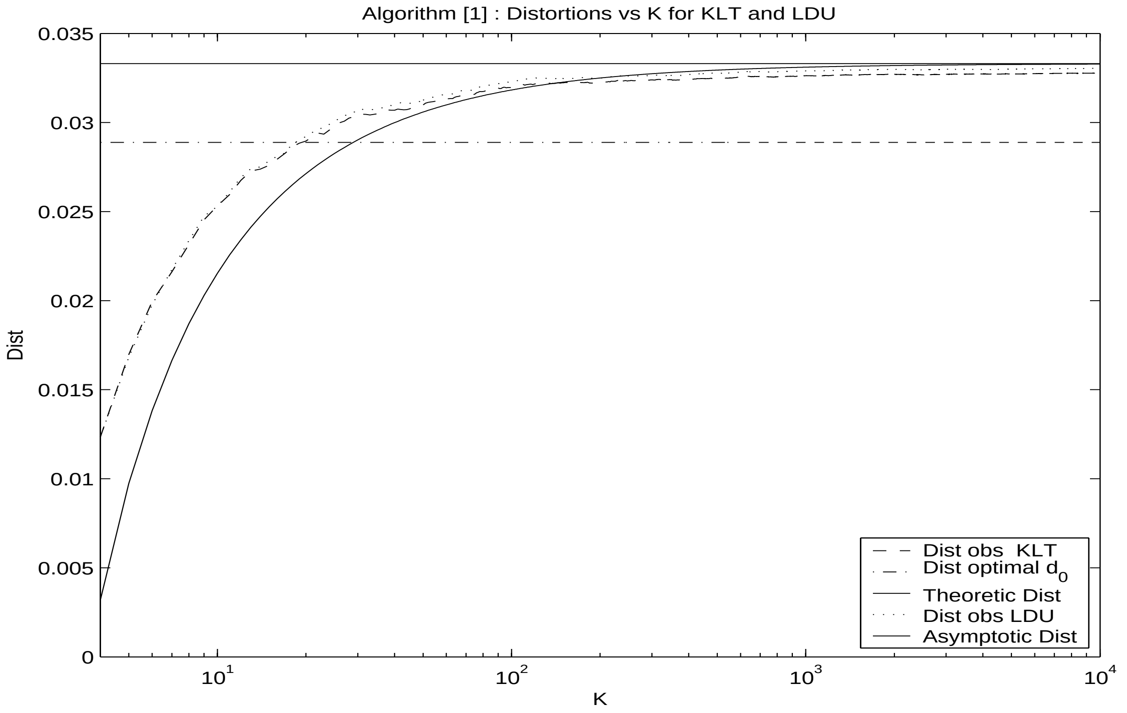

Algorithm [1] : Distortions vs K for KLT and LDU

Figure 4.8: Distortions for algorithm $[\mathit{1}]$ *vs* K, $r_0 = 2.3$ b/s.

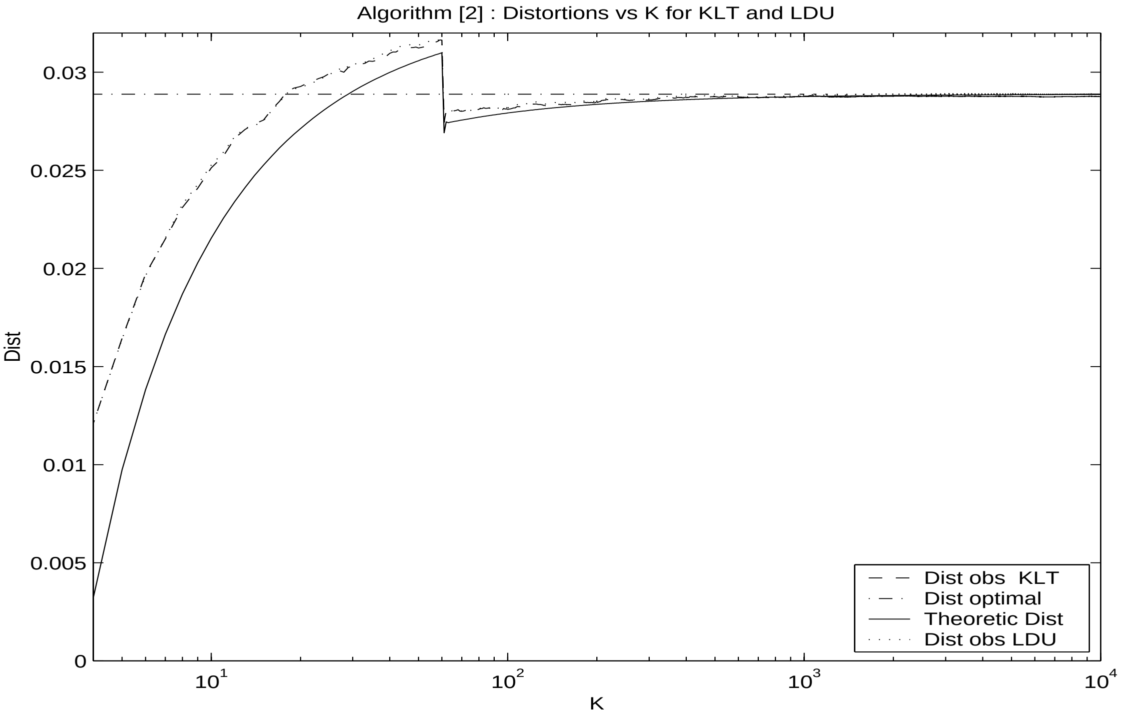

Algorithm [2] : Distortions vs K for KLT and LDU

Figure 4.9: Distortions for algorithm $[\mathit{2}]$ *vs* K, $r_0 = 2.3$ b/s.

4.8 Conclusions

Summarizing the framework, for sufficiently high rate, a particular point of the rate-distortion of a source with given covariance is chosen by fixing either the desired rate, or distortion. For transform coders using ECUQ, this is equivalent to choose a particular quantization stepsize. In backward adaptive transform coding, the decoder has *a priori* no knowledge about the statistics of the source. Thus, the tracking without side information must depend only on the previously decoded data.

Numerical results show that backward adaptive systems designed with a constant stepsize should converge to the target rate-distortion point for both the unitary and the causal approaches. In the latter case, the effects of the noise feedback caused by the closed loop implementation are noticeable at low rates only (below approximately 2 b/s).

Systems with fixed stepsizes may however result in uncontrolable variations of the actual rate-distortion performance if the source statistics change. These variations may be accounted for by using algorithms for which both the quantization stepsize and the transform are backward adaptive. In order to model the statistical behaviour of these systems, we assumed a stationary Gaussian vectorial source, whose covariance matrix is unknown. We showed that an algorithm using a Sheppard's correction on the estimate of the covariance matrix allows one to reach the target rate-distortion point; without this correction, there is a mismatch between the actual and the target rate-distortion performance of the system. These mismatches vanish for high resolution systems (small distortion/high rates). The proposed models match accurately the convergence process for both algorithms and transforms at rates higher than approximately 2.5 b/s.

4.A Perturbation of the determinant (4.15)

To simplify the notations, let us denote by $|.|$ the determinant $\det(.)$, let A be a square nonsingular matrix with elements $\{a_{ij}\}$, and $\widehat{A} = A + \delta A$, where δA is a perturbation matrix whose elements $\{\delta a_{ij}\}$ are small in comparison with the $\{a_{ij}\}$. In a first step, we compute $|A + \delta A|$ and take then care of the exponent $\frac{1}{N}$. One should now compute $|A + \delta A|$ which by the Taylor theorem may be approximated as

$$|A+\delta A| \approx |A| + \text{tr}\{(\delta A)^T \frac{\partial |A|}{\partial A}\} + \frac{1}{2}\sum_{i,j} \delta a_{ij}\, \text{tr}\{\frac{\partial^2 |A|}{\partial a_{ij}\partial A}\delta A^T\}. \tag{4.38}$$

The following properties [78] are now necessary to compute the second and third terms of (4.38). Denoting by A_{ij} the cofactor of a_{ij},

$$\begin{array}{lllll}
(a) & \frac{\partial |A|}{\partial a_{ij}} & & = & A_{ij} \\
(b) & \frac{\partial \ln |A|}{\partial A} & & = & (A^T)^{-1} = \frac{1}{|A|}\frac{\partial |A|}{\partial A} \\
(c) & A^T(A^T)^{-1} = I & \Rightarrow & & \frac{\partial A^T}{\partial a_{ij}}(A^T)^{-1} + A^T\frac{\partial (A^T)^{-1}}{\partial a_{ij}} = 0 \Rightarrow \frac{\partial (A^T)^{-1}}{\partial a_{ij}} = -(A^T)^{-1}\frac{\partial A^T}{\partial a_{ij}}(A^T)^{-1}
\end{array} \tag{4.39}$$

The second term of expression (4.38) may now, using prop. (b), be written as

$$\text{tr}\{(\partial A)^T \frac{\partial |A|}{\partial A}\} = \text{tr}\{\partial A^T |A| (A^T)^{-1}\}. \tag{4.40}$$

To compute the third term of (4.38), let us rewrite

$$\begin{aligned}
\frac{\partial}{\partial a_{ij}}\left(\frac{\partial |A|}{\partial A}\right) &= \frac{\partial(|A|(A^T)^{-1})}{\partial a_{ij}} \\
&= \underbrace{\frac{\partial |A|}{\partial a_{ij}}}_{A_{ij}}(A^T)^{-1} + \underbrace{|A|\frac{\partial (A^T)^{-1}}{\partial a_{ij}}}_{-(A^T)^{-1}\frac{\partial A^T}{\partial a_{ij}}(A^T)^{-1}} \\
&= A_{ij}(A^T)^{-1} - |A|(A^T)^{-1}\frac{\partial A^T}{\partial a_{ij}}(A^T)^{-1}.
\end{aligned} \tag{4.41}$$

The third term be then be written as

$$\begin{aligned}
\tfrac{1}{2}\sum_{i,j}\delta a_{ij}\,\text{tr}\{\frac{\partial^2|A|}{\partial a_{ij}\partial A}\delta A^T\} &= \tfrac{1}{2}\sum_{i,j}\delta a_{ij}\,\text{tr}\{A_{ij}(A^T)^{-1}\delta A^T\} - \frac{1}{2}\sum_{i,j}\delta a_{ij}\,\text{tr}\{|A|(A^T)^{-1}\frac{\partial A^T}{\partial a_{ij}}(A^T)^{-1}\delta A^T\} \\
&= \tfrac{1}{2}\sum_{i,j}\delta a_{ij}\underbrace{A_{ij}}_{\frac{\partial |A|}{\partial a_{ij}} = \left(\frac{\partial |A|}{\partial A}\right)_{ij} = |A|(A^T)^{-1}_{ij}}\text{tr}\{(A^T)^{-1}\delta A^T\} - \frac{|A|}{2}\sum_{i,j}\delta a_{ij} tr\{(A^T)^{-1}\frac{\partial A^T}{\partial a_{ij}}(A^T)^{-1}\delta A^T\} \\
&= \tfrac{|A|}{2}\sum_{i,j}\delta a_{ij}(A^T)^{-1}_{i,j}\,\text{tr}\{(A^T)^{-1}\delta A^T\} - \frac{|A|}{2}\text{tr}\{(A^T)^{-1}\delta A^T(A^T)^{-1}\delta A^T\} \\
&= \tfrac{|A|}{2}\left(\text{tr}\{\delta A A^{-1}\}\right)^2 - \tfrac{|A|}{2}\text{tr}\{\delta A(A)^{-1}\delta A(A)^{-1}\}.
\end{aligned} \tag{4.42}$$

Hence, eq. (4.38) may be approximated as

$$|A+\delta A| \approx |A|\left(1 + \underbrace{\text{tr}\{\delta A A^{-1}\}}_{\alpha} + \underbrace{\frac{1}{2}\left[(\text{tr}\{\delta A A^{-1}\})^2 - \text{tr}\{\delta A(A)^{-1}\delta A(A)^{-1}\}\right]}_{\beta}\right). \tag{4.43}$$

Denoting by α and β the underbraced terms of (4.38), we obtain now for $|A+\delta A|^{\frac{1}{N}}$

$$\begin{aligned} |A+\delta A|^{\frac{1}{N}} &\approx |A|^{\frac{1}{N}}\left[1+\tfrac{1}{N}(\alpha+\beta)+\tfrac{1}{2N}(\tfrac{1}{N}-1)(\alpha+\beta)^2\right] \\ &\approx |A|^{\frac{1}{N}}\left[1+\tfrac{1}{N}\alpha+\tfrac{1}{N}\beta+\tfrac{1-N}{2N^2}(\alpha^2+2\alpha\beta+\beta^2)\right], \end{aligned} \tag{4.44}$$

where we neglect the terms $\alpha\beta$ and β^2 for which δA is involved at a power superior than 2. We obtain

$$\begin{aligned} |A+\delta A|^{\frac{1}{N}} \approx\; & |A|^{\frac{1}{N}} \\ & \times\left[1+\tfrac{1}{N}\,\mathrm{tr}\,\{\delta A A^{-1}\}+(\tfrac{1}{2N})\big((\,\mathrm{tr}\,\{\delta A A^{-1}\})^2-\,\mathrm{tr}\,\{\delta A(A)^{-1}\delta A(A)^{-1}\}\big)+\tfrac{1-N}{2N^2}\big(\,\mathrm{tr}\,\{\delta A A^{-1}\}\big)^2\right] \\ \approx\; & |A|^{\frac{1}{N}}\left[1+\tfrac{1}{N}\{\,\mathrm{tr}\,\delta A A^{-1}\}+\tfrac{1}{2N^2}\big(\,\mathrm{tr}\,\{\delta A A^{-1}\}\big)^2-\tfrac{1}{2N}\,\mathrm{tr}\,\{\delta A A^{-1}\delta A A^{-1}\}\right]. \end{aligned} \tag{4.45}$$

Setting $A=R$, $\delta A=\Delta R_K^{[2]}$, and taking expectation of (4.45) establishes the expression (4.11).

4.B Derivation of (4.22)

The three terms involving expectations in (4.15) will be computed separately. The following properties will be used, see for example [78]:

$$(A\otimes B)(C\otimes D)=AC\otimes BD, \tag{4.46}$$

$$\mathrm{tr}\,\{AB\}=\mathrm{vec}^T(B^T)\mathrm{vec}(A), \tag{4.47}$$

$$\mathrm{tr}\,\{ABC\}=\mathrm{vec}^T(A^T)(C\otimes I)\mathrm{vec}(B), \tag{4.48}$$

$$\mathrm{tr}\,\{A\otimes B\}=\,\mathrm{tr}\,\{A\}\,\mathrm{tr}\,\{B\}, \tag{4.49}$$

$$\mathrm{vec}(ABC)=(C^T\otimes A)\mathrm{vec}(B)=(I\otimes AB)\mathrm{vec}C. \tag{4.50}$$

- First term of (4.15): Using the definition of $\Delta R_{K,det}$ and $\Delta R_{K,sto}$ of (4.13) and the statistics of the ”one-shot” estimates of section 3.A, this term becomes

$$\begin{aligned} \tfrac{1}{N}\,\mathrm{E}\;\mathrm{tr}\,\{\Delta R_K^{[2]}R^{-1}\} &= \tfrac{1}{N}\left(\,\mathrm{tr}\,\{\Delta R_{K,det}^{[2]}R^{-1}\}+\underbrace{\mathrm{E}\;\mathrm{tr}\,\{\Delta R_{K,sto}^{[2]}R^{-1}\}}_{0}\right) \\ &\approx \tfrac{1}{N}\left[\tfrac{1}{K}\Sigma_{tot}-D^{[2]}(K)\right]\;\mathrm{tr}\,\{R^{-1}\}, \end{aligned} \tag{4.51}$$

with $\Sigma_{tot}=\displaystyle\sum_{i=N+1}^{N_1}D^{[1]}(i)+\sum_{i=N_1+1}^{K}D^{[2]}(i)$.

- Second term of (4.15): This is

$$\begin{aligned}
\tfrac{1}{2N^2}\,\mathrm{E}\,(\,\mathrm{tr}\,\{\Delta R_K^{[2]} R^{-1}\})^2 &= \tfrac{1}{2N^2}\,\mathrm{E}\,(\,\mathrm{tr}\,\{R^{-\frac{1}{2}}\Delta R_K^{[2]} R^{-\frac{1}{2}}\})^2 \\
&\approx \tfrac{1}{2N^2}(\mathrm{vec}^T R^{-\frac{1}{2}})(R^{-\frac{1}{2}} \otimes I)\underbrace{\mathrm{E}\,\mathrm{vec}\Delta R_K^{[2]}\mathrm{vec}^T\Delta R_K^{[2]}}(R^{-\frac{1}{2}} \otimes I)\mathrm{vec} R^{-\frac{1}{2}} \\
&\qquad \mathrm{E}\,\left(\mathrm{vec}\Delta R_{K,det}^{[2]} + \mathrm{vec}\Delta R_{K,sto}^{[2]}\right)\left(\mathrm{vec}^T\Delta R_{K,det}^{[2]} + \mathrm{vec}^T\Delta R_{K,sto}^{[2]}\right) \\
&\qquad \underbrace{\mathrm{vec}\Delta R_{K,det}^{[2]}\mathrm{vec}^T\Delta R_{K,det}^{[2]}}_{a_1} + 0 + 0 + \underbrace{\mathrm{E}\,\mathrm{vec}\Delta R_{K,sto}^{[2]}\mathrm{vec}^T\Delta R_{K,sto}^{[2]}}_{b_1}
\end{aligned} \tag{4.52}$$

We first compute separately the terms named a_1 and b_1.

– Term a_1: Since $\Delta R_{K,det}^{[2]}$ is a diagonal matrix, its contribution is the square of that involved in the first term of (4.51), weighted by $\frac{1}{2N^2}$ instead of $\frac{1}{N}$.

– Term b_1: Assuming small perturbation due to quantization and estimation noise, expanding the term $\Delta R_{K,sto}^{[2]}$ gives, considering the estimates $\Delta R^{[it1,2],q}$ as independent

$$\begin{aligned}
b_1 &= \mathrm{E}\,\mathrm{vec}\Delta R_{K,sto}^{[2]}\mathrm{vec}^T\Delta R_{K,sto}^{[2]} \\
&\approx \frac{1}{2N^2}\left(\underbrace{\sum_{i=1}^{N}\mathrm{vec}\Delta R_i^{(1)}\mathrm{vec}^T\Delta R_i^{(1)}}_{a} + \underbrace{\sum_{i=N+1}^{N_1}\mathrm{vec}\Delta R_i^{q[1](1)}\mathrm{vec}^T\Delta R_i^{q[1](1)}}_{b} + \underbrace{\sum_{i=N_1+1}^{K}\mathrm{vec}\Delta R_i^{q[2](1)}\mathrm{vec}^T\Delta R_i^{q[2](1)}}_{c}\right),
\end{aligned} \tag{4.53}$$

where the term $\mathrm{E}\,(\delta D_K^{[2]} I)^2 \mathrm{vec}(I)(\mathrm{vec} I)^T$ is neglected for the purpose if this first order perturbation analysis, at high resolution, and assuming sufficiently high K. Applying now the result (4.17) to the one shot estimates of their corresponding matrices yields

$$\begin{aligned}
a &= \sum_{i=1}^{N}\mathrm{vec}\Delta R_i^{(1)}\mathrm{vec}^T\Delta R_i^{(1)} \\
&\approx 2NR \otimes R \\
b &= \sum_{i=N+1}^{N_1}\mathrm{vec}\Delta R_i^{q[1](1)}\mathrm{vec}^T\Delta R_i^{q[1](1)} \\
&\approx 2(N_1 - N)R \otimes R + 2\sum_{i=N+1}^{N_1} D_i^{[1]}(R \otimes I + I \otimes R) \\
c &= \sum_{i=N_1+1}^{K}\mathrm{vec}\Delta R_i^{q[2](1)}\mathrm{vec}^T\Delta R_i^{q[2](1)} \\
&\approx 2(K - N_1)R \otimes R + 2\sum_{i=N+1+1}^{K} D_i^{[2]}(R \otimes I + I \otimes R)
\end{aligned} \tag{4.54}$$

Hence

$$\frac{1}{K^2}(a + b + c) \approx \frac{1}{K^2}\left[\underbrace{2KR \otimes R}_{A} + \left(\underbrace{R \otimes I}_{B} + \underbrace{I \otimes R}_{C}\right)(2\Sigma_{tot})\right], \tag{4.55}$$

and using the underbraced terms A, B and C, the term b_1 may be written as

$$b_1 \approx \frac{1}{2K^2N^2}(\text{vec}^T R^{-\frac{1}{2}})(R^{-\frac{1}{2}} \otimes I)(A+B+C)(R^{-\frac{1}{2}} \otimes I)\text{vec}R^{-\frac{1}{2}}. \tag{4.56}$$

- $*$ Contribution of A

$$\begin{aligned}
&\tfrac{2K}{2K^2N^2}(\text{vec}^T R^{-\frac{1}{2}})\,\underbrace{(R^{-\frac{1}{2}} \otimes I)\,\underbrace{(R \otimes R)(R^{-\frac{1}{2}} \otimes I)}_{R^{-\frac{1}{2}} \otimes R}}_{I \otimes R}\text{vec}R^{-\frac{1}{2}} \\
&= \tfrac{1}{KN^2}\text{vec}^T R^{-\frac{1}{2}}(I \otimes R)\text{vec}R^{-\frac{1}{2}} \\
&= \tfrac{1}{KN^2}\,\text{tr}\{R^{-\frac{1}{2}} R R^{-\frac{1}{2}}\} \\
&= \tfrac{1}{KN}
\end{aligned} \tag{4.57}$$

- $*$ Contribution of B

$$\begin{aligned}
&\tfrac{2\Sigma_{tot}}{2K^2N^2}\text{vec}^T R^{-\frac{1}{2}}(R^{-\frac{1}{2}} \otimes I)(R \otimes I)(R^{-\frac{1}{2}} \otimes I)\text{vec}(R^{-\frac{1}{2}}) \\
&= \tfrac{\Sigma_{tot}}{K^2N^2}\text{vec}^T R^{-\frac{1}{2}} I_{N^2}\text{vec}(R^{-\frac{1}{2}} \\
&= \tfrac{\Sigma_{tot}}{K^2N^2}\,\text{tr}\{R^{-1}\},
\end{aligned} \tag{4.58}$$

where I_{N^2} denotes the $N \times N$ Identity matrix.

- $*$ Contribution of C

$$\begin{aligned}
&\tfrac{\Sigma_{tot}}{K^2N^2}\text{vec}^T R^{-\frac{1}{2}}(R^{-\frac{1}{2}} \otimes I)(I \otimes R)(R^{-\frac{1}{2}} \otimes I)\text{vec}(R^{-\frac{1}{2}}) \\
&= \tfrac{\Sigma_{tot}}{K^2N^2}\,\text{tr}\{R^{-1}\},
\end{aligned} \tag{4.59}$$

which is the same contribution as B.

The term b_1 may thus be approximated as

$$b1 \approx \frac{1}{KN}\left[1 + \frac{2\Sigma_{tot}\,\text{tr}\{R^{-1}\}}{N}\right]. \tag{4.60}$$

Grouping the terms a_1 and b_1 yields

$$\frac{1}{2N^2}\,\text{E}\,(\,\text{tr}\{\Delta R_K^{[2]} R^{-1}\})^2 \approx \frac{1}{KN} + \text{tr}\{R^{-1}\}(\frac{2\Sigma_{tot}}{K^2N^2}) + \frac{\left(\text{tr}\{R^{-1}\}\right)^2}{2N^2}\left[\left(\frac{\Sigma_{tot}}{K} - D^{[2]}(K)\right)^2\right], \tag{4.61}$$

which is the expression (4.20).

- Third term of (4.15): Let G be $R^{-\frac{1}{2}}\Delta R_K^{[2]}R^{-\frac{1}{2}}$. Then we have vec$G = (R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})vec\Delta R_K^{[2]}$, and we get

$$\begin{aligned}
-\tfrac{1}{2N}\,\mathrm{E}\,tr\{\Delta R_K^{[2]}R^{-1}\Delta R_K^{[2]}R^{-1}\} &= \ \mathrm{E} - \tfrac{1}{2N}\,\mathrm{tr}\,\{GG\} \\
&= -\tfrac{1}{2N}\,\mathrm{E}\ \mathrm{tr}\,\{\mathrm{vec}G\mathrm{vec}^T G\} \\
&= -\tfrac{1}{2N}\,\mathrm{E}\ \mathrm{tr}\,\{(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})\underbrace{E\mathrm{vec}\Delta R_K^{[2]}\mathrm{vec}^T\Delta R_K^{[2]}}_{\underbrace{E\mathrm{vec}\Delta R_{K,det}^{[2]}\mathrm{vec}^T\Delta R_{K,det}^{[2]}}_{a_2} + \underbrace{E\mathrm{vec}\Delta R_{K,sto}^{[2]}\mathrm{vec}^T\Delta R_{K,sto}^{[2]}}_{b_2}}(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})\} \\
&= -\tfrac{1}{2N}\,\mathrm{E}\ \mathrm{tr}\,\{(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})a_2(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})\} \\
&\quad -\tfrac{1}{2N}\,\mathrm{E}\ \mathrm{tr}\,\{(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})b_2(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})\}.
\end{aligned} \tag{4.62}$$

The contribution of these terms is computed separately:

  - Contribution of the first term in (4.62)

$$\begin{aligned}
&-\tfrac{1}{2N}\,\mathrm{E}\ \mathrm{tr}\,\{(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})a_2(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}}) \\
&\approx \tfrac{1}{2N}(\tfrac{1}{K}\Sigma_{tot} - D_K^{[2]})^2\ \underbrace{\mathrm{tr}\,\{(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})}_{\mathrm{tr}\,\{R^{-1}\otimes R^{-1}\}\approx(\,\mathrm{tr}\,\{R^{-1}\})^2}\}.
\end{aligned} \tag{4.63}$$

  - Contribution of the second term in (4.62)

$$\begin{aligned}
&-\tfrac{1}{2N}\,\mathrm{E}\ \mathrm{tr}\,\{(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})b_2(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}}) \\
&= -\tfrac{1}{2N}\,\mathrm{tr}\,\{(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})\tfrac{1}{K^2}\left[A + (B + C)2\Sigma_{tot}\right](R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})\},
\end{aligned} \tag{4.64}$$

  where the terms A, B and C have been define in (4.55). Their respective contribution are

$$\begin{aligned}
A\ &:\ -\tfrac{1}{NK}\,\mathrm{tr}\,\{(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})(R \otimes R)(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})\} = -\tfrac{1}{NK}\,\mathrm{tr}\,\{\underbrace{(R \otimes R)^{-\frac{1}{2}}(R \otimes R)^{\frac{1}{2}}}_{I_{N^2}}\} = -\tfrac{N}{K}, \\
B\ &:\ -\tfrac{1}{NK^2}\Sigma_{tot}\,\mathrm{tr}\,\{\underbrace{(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})(R \otimes I)(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})}_{\mathrm{tr}\,\{I\otimes R^{-1}=N\,\mathrm{tr}\,R^{-1}\}}\} = -\tfrac{\Sigma_{tot}\,\mathrm{tr}\,R^{-1}}{K^2}, \\
C\ &:\ -\tfrac{1}{NK^2}\Sigma_{tot}\,\mathrm{tr}\,\{(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})(I \otimes R)(R^{-\frac{1}{2}} \otimes R^{-\frac{1}{2}})\} = -\tfrac{\Sigma_{tot}\,\mathrm{tr}\,R^{-1}}{K^2}.
\end{aligned} \tag{4.65}$$

By regrouping the contributions of the terms in (4.63) and (4.65) we get for the third term in (4.15)

$$-\frac{1}{2N}\,\mathrm{E}\ \mathrm{tr}\,\{\Delta R_K^{[2]}R^{-1}\Delta R_K^{[2]}R^{-1} \approx \frac{N}{K} - \frac{2\Sigma_{tot}\,\mathrm{tr}\,\{R^{-1}\}}{K^2} - \frac{(\,\mathrm{tr}\,\{R^{-1}\})^2}{2N}\left[(\frac{1}{K}\Sigma_{tot} - D_K^{[2]})^2\right]. \tag{4.66}$$

Grouping finally the terms of (4.51), (4.61) and (4.66), and by retaining only the first order perturbation terms (linear in D) yields the distortion (4.22).

Chapter 5

Generalized MIMO Prediction

For vectorial sources presenting memory, we show in this chapter that the optimal causal decorrelating scheme can be described by means of a prediction matrix whose entries are optimal prediction filters. This decorrelating procedure leads to the notion of "generalized MIMO prediction", in which a certain degree of non causality may be allowed for the off-diagonal prediction filters. In the case of non causal intersignal filters, the optimal MIMO predictor is still lower triangular, and hence "causal", in a wider sense. The notion of causality is generalized in the sense that causality between channels becomes processing the channels in a certain order. We then show that two previously introduced transformations, in the context of subband coding, appear as special cases of this generalized MIMO prediction. As the previously described causal LDU transform, realistic coding implementations of the latter two approaches should involve closed loop structures for the prediction. We show that though these approaches are equivalent in the limit of high rates, triangular MIMO prediction may be more efficient than its classical counterpart. In this case, we show that the optimal ordering of the scalar signals (w.r.t. the coding performance at high rate) corresponds to the case where they get decorrelated by order of decreasing variance. Finally, we present some applications of these results to wideband speech coding.

5.1 Introduction

In the transform coding framework, previous chapters showed that the optimal causal transform is a lower triangular and unit diagonal matrix, which corresponds to a (Lower-Diagonal-Upper) factorization of the autocorrelation matrix of the signal. The rows of this matrix are optimal prediction filters for the corresponding component of the vector to be coded, and the transformed coefficients are the optimal prediction errors. As in classical (A)DPCM, the prediction should be implemented in closed loop around the quantizers, that is, using the previously quantized samples. As in (A)DPCM also, we showed that a quantization noise feedback occurs for which closed form expressions can be obtained. In this chapter, we apply this causal decorrelation approach to the optimal coding of vectorial signals, as for example those obtained by subband filtering stereo, or multichannel audio signals. In this case, the vectorial source $\underline{x}$ may present both temporal redundancies (between the samples $\underline{x}_k$ at different time instants) and spatial redundancies (between the scalar sources x_i). Thus, instantaneous decorrelation such as that performed by a decorrelating matrix (KLT or LDU) is not optimal, even for Gaussian sources; further decorrelation may be achieved by exploiting the temporal correlation structure of the vectorial source. Optimal coding of vectorial signals will refers to decorrelating strategies which remove both spatial and temporal dependencies; the source model is Gaussian. As in the analysis of chapter 2, the coding gain G_T will be the criterion of merit which allows one to evaluate the coding performance of a transformation T. It corresponds to the factor by which the distortion is reduced because of T,

$$G_T = \frac{\mathrm{E}\,||\underline{\widetilde{x}}||_I^2}{\mathrm{E}\,||\underline{\widetilde{y}}||_T^2}. \tag{5.1}$$

For the causal decorrelating approach introduced at the begining of this thesis, the analysis of the coding is again made for two cases: neglecting the effects of the noise feedback in a first step, and accounting for them in a second step.

By considering vectors of infinite size, we show in section 5.2 that one can get frequential expressions for the coding gains. In this case, the optimal causal decorrelating scheme can be described by means of a prediction matrix whose entries are optimal prediction filters. The diagonal filters are scalar intrasignal prediction filters. The off-diagonal predictors are Wiener filters performing the intersignal decorrelation.

This decorrelating procedure leads in section 5.3 to the notion of ”generalized MIMO prediction”, in which a certain degree of non causality may be allowed for the off-diagonal prediction filters. In the case of non causal intersignal filters, the optimal MIMO predictor is still lower triangular, and hence “causal”, in a wider sense. The notion of causality is generalized : the causality between channels becomes processing the channels in a certain order. Some signals may be coded using the coded/decoded versions of the “previous” signals. We also show in section 5.3 that two previously introduced decorrelation approaches are actually special cases of this so-called generalized MIMO prediction.

An interesting (and empiricial) result of chapter 2 is that if the quantization noise feedback is taken into account, the efficiency of the interband decorrelation depends on the order in which the decorrelation between the signals is processed. We present in the fourth section of this chapter a new theorem concerning the optimal ordering of the signals for a triangular “causal” MIMO predictor, namely the ordering which minimizes the quantization noise feedback.

The fifth part of this chapter deals with optimal triangular MIMO prediction with finite prediction orders. Despite the non causality in the classical sense of this approach, the optimal triangular MIMO prediction is well suited for frame based audio coding, which allows a certain degree of non causality in the coding procedure. When FIR filters are used to perform the intersignal decorrelation, we will show that the optimal positioning of a finite number of taps is fairly straightforward.

An application of the proposed coding procedure is presented in the framework of wideband speech coding in section 5.5. Finally, the last section summarizes the results of this chapter and draws some conclusions.

5.2 Optimal Causal Coding of Vectorial Signals

Let us consider a Gaussian vector source $\underline{x}$ with covariance matrix $R_{\underline{xx}}$. Each sample vector $\underline{x}_k$ of $\underline{x}$ may be transformed by means of an optimal causal transform L, and the resulting transform y_i components scalar quantized and further entropy coded. In this framework, the optimal causal transform is of the form

$$L = \begin{bmatrix} 1 & & & \\ \star & \ddots & \mathbf{0} & \\ \vdots & \ddots & \ddots & \\ \star & \cdots & \star & 1 \end{bmatrix},$$

where the $\star$ represent optimal prediction coefficients. In other words, L is such that

$$L R_{\underline{xx}} L^T = R_{\underline{yy}} = \overline{\text{diag}}\{\sigma^2_{y_1} \cdots \sigma^2_{y_N}\}, \tag{5.2}$$

where $diag\{\underline{a}\}$ represent the diagonal matrix with diagonal $\underline{a}$. Since each prediction error y_i is orthogonal to the subspaces generated by the $\underline{x}_{1:i-1}$, the transform coefficients y_i are orthogonal, and $R_{\underline{yy}}$ is diagonal. It follows that

$$R_{\underline{xx}} = L^{-1} R_{\underline{yy}} L^{-T}, \tag{5.3}$$

which represents the LDU factorization of $R_{\underline{xx}}$.

5.2.1 Case of Negligible Feedback

Let us now consider the case in which each vector $\underline{X}_k$ to be coded is composed of a succession of samples of a stationary vectorial signal $\underline{x}_k = [x_{1,k} \cdots x_{M,k}]^T$, $\underline{X}_k = [\underline{x}_0^T \ \underline{x}_1^T \cdots \underline{x}_k^T]^T$. The transform vector $\underline{Y}_k = L\underline{X}_k = [\underline{y}_0^T \ \underline{y}_1^T \cdots \underline{y}_k^T]^T$ with $\underline{y}_k = [y_{1,k} \cdots y_{M,k}]^T$. For these vectorial signals, it is interesting to consider the limiting case in which the dimension k goes to infinity. In this case, the optimal transform L will lead to a signal $\underline{y}_k$, asymptotically stationary too, since L will become block Toeplitz (with blocks of size $M \times M$). In this case, the coding gain (5.1) becomes

$$G_L^{(0)} = \lim_{k\to\infty} \left(\frac{\det\left[\text{diag}\left(R_{\underline{X}_k\underline{X}_k}\right)\right]}{\det\left[\text{diag}\left(L R_{\underline{X}_k\underline{X}_k} L^T\right)\right]} \right)^{\frac{1}{Mk}} = \left(\frac{\det\left[\text{diag}\left(R_{\underline{x}_k\underline{x}_k}\right)\right]}{\det[\text{diag}\left(R_{\underline{y}_k\underline{y}_k}\right)]} \right)^{\frac{1}{M}} = \left(\frac{\prod_{i=1}^{M} \sigma_{x_i}^2}{\prod_{i=1}^{M} \sigma_{y_i}^2} \right)^{\frac{1}{M}} \tag{5.4}$$

where $y_{i,k}$ is the optimal prediction error of infinite order of $x_{i,k}$, based on $\{\underline{x}_{1:N,-\infty:k-1}, \underline{x}_{1:i-1,k}\}$[1]. We shall continue to denote by L_i (now of infinite dimension) the vector of the corresponding prediction coefficients.

There exists a frequency domain expression for $\prod_{i=1}^{M} \sigma_{y_i}^2$. Since $\underline{y}$ is a totally decorrelated signal, its power spectral density matrix can be written as

$$S_{\underline{yy}}(f) = R_{\underline{yy}} = \overline{\text{diag}}\,\{\sigma_{y_1}^2, \ldots, \sigma_{y_M}^2\}. \tag{5.5}$$

If we now describe the prediction operation in the frequency domain, the prediction error should be written as $\underline{Y}(f) = L(f)\,\underline{X}(f)$, where $\underline{Y}(f)$ and $\underline{X}(f)$ denote the Fourier transforms of $\underline{y}_k$ and $\underline{x}_k$. The $M \times M$ matrix $L(f)$ denotes the Fourier transform of the prediction error filter. Hence we have

$$S_{\underline{yy}} = L(f)\, S_{\underline{xx}}(f)\, L^H(-f), \tag{5.6}$$

where ^{H} denotes the Hermitian transposition, and $L^H(-f) = L^T(f)$ since we consider real signals. Thus, we can write

$$\begin{aligned} \prod_{i=1}^{M} \sigma_{y_i}^2 &= e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln[\det(S_{\underline{yy}}(f))] df} \\ &= e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \{\ln[\det(S_{\underline{xx}}(f))] + 2\ln[\det(L(f))]\} df} \\ &= e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln[\det(S_{\underline{xx}}(f))]\, df} \end{aligned} \tag{5.7}$$

where we used the following property (due to the monic and causal diagonal prediction filters, see Appendix 5.A)

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\det[L(f)]\, df = 0. \tag{5.8}$$

The coding gain with negligible feedback (or infinite resolution) is thus

$$G_L^{(0)} = \left(\frac{\prod_{i=1}^{M} \sigma_{x_i}^2}{e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln[\det(S_{\underline{xx}}(f))]\, df}} \right)^{\frac{1}{M}}. \tag{5.9}$$

[1] The notation $\underline{x}_{i:j,k:K}$ denotes the set of samples of the components $x_i, x_{i+1} \cdots x_{j-1}, x_j$ of $\underline{x}$ at instants $k, k+1 \cdots K-1, K$.

5.2.2 Noise Feedback Effects on the Coding Gain

If we now consider the effects of the quantization in the closed loop an analysis similar to that of 2.5.2 can be made. The gain $G_L^{(1)}$ can then be expressed as

$$G_L^{(1)} \approx \lim_{k\to\infty} \left(\frac{\det[\,\text{diag}\,(R_{\underline{X}_k\underline{X}_k})]}{\det[\,\text{diag}\,(LR_{\underline{X}_k\underline{X}_k}L^T + \sigma_q^2\overline{L}\,\overline{L}^T)]} \right)^{\frac{1}{Mk}} \approx \left(\frac{\prod_{i=1}^{M} \sigma_{x_i}^2}{\prod_{i=1}^{M} [\sigma_{y_i}^2 + \sigma_q^2(\|L_i\|^2 - 1)]} \right)^{\frac{1}{M}} , \tag{5.10}$$

which leads to

$$G_L^{(1)} \approx G_L^{(0)} \left(1 - \sigma_q^2 \frac{1}{M} \sum_{i=1}^{M} \frac{\|L_i\|^2 - 1}{\sigma_{y_i}^2} \right) , \tag{5.11}$$

where L and $\sigma_{y_i}^2$ refer to non perturbed quantities.

As in the ideal case, one can derive an expression for $G_L^{(1)}$ in the frequency domain. Under the high resolution assumption, the quantization errors are decorrelated. Hence we can write $S_{\tilde{y}\tilde{y}}(f)$ as $\overline{\text{diag}}\,\{\sigma_{\tilde{y}_1}^2, \ldots, \sigma_{\tilde{y}_M}^2\}$, or, equivalently, as $\sigma_q^2\, I_M$, in the case of an optimal bit assignment. Using a similar analysis as in section 2.5.2, the coding gain taking into account the perturbation effects up to first order may be written as (see 5.B)

$$G_L^{(1)} \approx G_L^{(0)} \left[1 + \frac{\sigma_q^2}{M} \left(- \int_{-\frac{1}{2}}^{\frac{1}{2}} \text{tr}\, \left(S_{\underline{xx}}^{-1}(f) \right) df + \sum_{i=1}^{M} \frac{1}{\sigma_{y_i}^2} \right) \right] \tag{5.12}$$

where, comparing with equation (5.11), the term $\int_{-\frac{1}{2}}^{\frac{1}{2}} \text{tr}\,\{S_{\underline{xx}}^{-1}(f)\}df$ corresponds to $\sum_{i=1}^{M} \frac{\|L_i\|^2}{\sigma_{y_i}^2}$. Thus, maximizing $G^(1)_L$ entails maximizing the sum of the inverses of the prediction error variances. This results was obtained in the transform coding framework of chapter 2 also. Empirical evidence was given that maximizing $G^(1)_L$ entails processing the signals by order of decreasing variance. This will be proven in section 5.4.

5.3 Linear Prediction of Subband Signals

In the case of subband coding, in which the components x_i of the vectorial signal $\underline{x}$ correspond to the subband signals, we will now show that two previously introduced transformations for maximizing the coding gain are special cases of a causal unit diagonal transformation. Moreover, the equivalence of these transforms in the ideal case (considering $G_{TC}^{(0)}$) is a consequence of the LDU nature of the optimal transformation.

5.3.1 Subband Coding

Subband coding schemes decompose a source data stream into a number of subsignals, each having a passband equal to a fraction of the bandwidth of the original signal, and the subsignals collectively cover the entire bandwidth of the original signal. Because a reduction in bandwidth of these subsignals, they are typically downsampled so that an efficient representation of the original source data may be obtained. A bit allocation procedure is then performed that assigns a bit rate to each subsignal subject to an overall bit rate constraint, and finally each subsignal is encoded *independently* of the other subsignals. Many subband systems use filterbanks, whose filters are designed to satisfy a perfect reconstruction property, see e.g. [79]. In the example of figure 5.1, the subband signals obtained from downsampling and filtering some process x are further independently coded. Perfect reconstruction is assumed in absence of quantization.

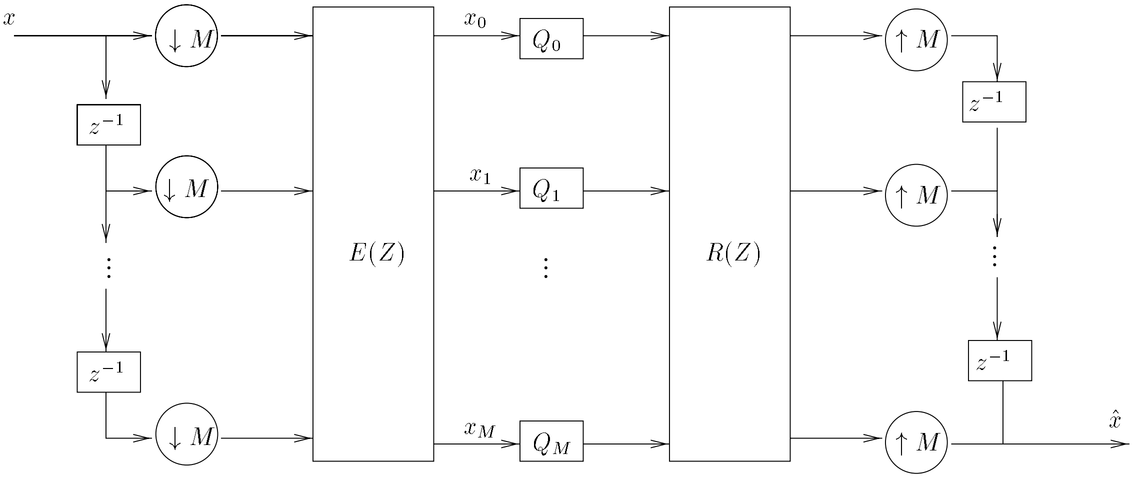


Figure 5.1: Polyphase representation of a filter bank.

Several results exist which describe the coding efficiency of this structure. For ideal filters with equal bandwidth (non overlapping brickwall frequency responses), the coding gain is asymptotically the same, w.r.t. the number of subbands, as that of transform coding[2]. This is not the case of other subband decompositions, which may be suboptimal even for infinite number of subbands (e.g. [80]). Various performance comparisons with TC and DPCM, and applications to images and audio coding may be found in [79, 14, 11].

5.3.2 Linear Prediction of Subband Signals

If now a prediction stage is applied to the subband signals before quantization, a question of interest is to know whether, for realizable filterbanks, this structure is optimal (in the sense that it totally decorrelates the input).

For finite prediction order on the one hand, the subband approach has been shown to be more efficient in the sense that it minimizes the combined prediction errors of the subbands w.r.t. that of the fullband for a given order[3]. This was shown in [81] for Gaussian signals and ideal analysis and synthesis filters. Similarly, for Gaussian AR sources, the pth-order entropy of the combined subbands is lower for subband signals than

[2] Note that the coding gain may decrease by increasing this number for subband coding, whereas this is not the case for TC [79].

[3] In this sense, the resulting $p - th$ order is called ”super-optimal” [81].

that of the fullband signal, for any finite $p.$ These results are confirmed for lossless coding of audio signals, QMF banks, finite linear prediction order and context based non linear prediction in [82]. (Interestingly, the problem of attributing the optimal order p_i subject to the constraint that $\sum_{i=1}^{N} = p$ resembles that of allocating the bits for a tranform coder, because the prediction error variances are non increasing variances of the prediction order, just as the distortions with the bits.)

For optimal prediction on the other hand, Fischer showed in [83] that, except for special cases[4], independently coding the subbands x_i instead of the fullband signal x is, from a rate distortion viewpoint, suboptimal. The analysis assumes realizable perfect reconstruction QMF filters [84], and wide sense stationary Gaussian processes. The result is drawn from the high resolution rate-distortion function of the system, which depends on the variances of the subband signals, which in turn can be analytically derived from the power spectrum densities $S_{x_i x_i}(f)$ obtained for the particular considered filters. The geometric mean of these variances is greater than the prediction error of the fullband signal because some interchannel correlation remain, which can not be further removed. Two approaches aimed of totally decorrelating the subband signals in order to maximize the coding gain where then proposed.

5.3.3 Two Causal Decorrelation Approaches Compared

On the one hand, Maison and Vanderdorpe [85] introduced in the classical subband coding scheme a matricial filtering transformation $T(z)$, which transforms the vectorial signal $\underline{x}_k = [x_{1,k} ... x_{M,k}]^T$ into the vectorial signal $\underline{y}_k = T(q)\,\underline{x}_k$ (where q is the unit delay operator). This approach corresponds to the causal MIMO prediction : $T(z) = \sum_{k=0}^{\infty} T_k\, z^{-k}$, where T_0 is lower triangular and unit diagonal. The MIMO predictor is assumed to be of infinite order. In order to keep the structure causal, each sample of the subband i is predicted by means of the past samples of all subbands, and by means of the present samples of lower index only. In the case $M = 2$, the MIMO predictor is made of two intraband scalar predictors and two interband scalar predictors. It was shown in [85] that such a transformation leads to an optimal coding gain $G_{TC}^{(0)}$, because the components of the resulting process are totally decorrelated. For AR(p) and finite prediction order, this approach may also outperform fullband prediction for some orders smaller than p [86].

On the other hand, Wong used the following triangular transform [87]: in the case $M = 2$,

$$\begin{aligned} T(z) &= \begin{bmatrix} 1 & 0 \\ 0 & T_{22}(z) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ W_{21}(z) & 1 \end{bmatrix} \begin{bmatrix} T_{11}(z) & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} T_{11}(z) & 0 \\ T_{22}(z)W_{21}(z)T_{11}(z) & T_{22}(z) \end{bmatrix} . \end{aligned} \tag{5.13}$$

[4] *e.g.* if the filters of $E(z)$ are ideal (brickwall), or if the p.s.d. $S_{xx}(f)$ is symmetric about $f = \frac{1}{4}$ in the case of two subbands.

The scalar prediction error filter $T_{11}(z)$ whitens $x_{1,k}$ and yields $y_{1,k}$, $W_{21}(z)$ is a (noncausal) Wiener filter estimating $x_{2,k}$ from $y_{1,k}$, and $T_{22}(z)$ whitens the resulting error signal to yield $y_{2,k}$. This transform hence uses only one interband predictor $W_{21}(z)$. The loss in degrees of freedom due to the loss of one interband predictor (T_{12} in Maison and Vandendorpe's transformation) is balanced by the non causality of this remaining unique interband predictor. Using a similar analysis as in [83], Wong showed that the suboptimality due to the non ideal subband filters vanishes, or equivalently, that the subband redundancy is removed from the coding procedure, assuming high rate and Gaussianity for all the signals to be quantized. We will now show that these two transformations can both be expressed as lower triangular unit diagonal transforms, simply by reorganizing the samples in the vector to be coded. Let us write these transformations in the case of two subbands, and for a finite frame of signal :

- In Maison and Vandendorpe's approach, $L\underline{X}_k$ can be witten as

$$L_1\underline{X}_k = \begin{bmatrix} 1 & & & & & & & \\ \star & \ddots & & & & \mathbf{0} & & \\ \vdots & \ddots & \ddots & & & & & \\ \star & \cdots & \star & 1 & & & & \\ \star & \cdots & \cdots & \star & 1 & & & \\ \vdots & & & \vdots & \star & \ddots & & \\ \vdots & & & \vdots & \vdots & \ddots & \ddots & \\ \star & \cdots & \cdots & \star & \star & \cdots & \star & 1 \end{bmatrix} \begin{bmatrix} x_{1,0} \\ x_{2,0} \\ -- \\ x_{1,1} \\ x_{2,1} \\ -- \\ \vdots \\ -- \\ x_{1,k} \\ x_{2,k} \end{bmatrix} = \begin{bmatrix} y_{1,0} \\ y_{2,0} \\ -- \\ y_{1,1} \\ y_{2,1} \\ -- \\ \vdots \\ -- \\ y_{1,k} \\ y_{2,k} \end{bmatrix} = \underline{Y}_k$$

- In Wong's approach, $L\mathcal{P}\underline{X}_k$ can be written as

$$L_2\mathcal{P}\underline{X}_k = \begin{bmatrix} 1 & & & & & & & \\ \star & \ddots & & & & \mathbf{0} & & \\ \vdots & \ddots & \ddots & & & & & \\ \star & \cdots & \star & 1 & & & & \\ \star & \cdots & \cdots & \star & 1 & & & \\ \vdots & & & \vdots & \star & \ddots & & \\ \vdots & & & \vdots & \vdots & \ddots & \ddots & \\ \star & \cdots & \cdots & \star & \star & \cdots & \star & 1 \end{bmatrix} \begin{bmatrix} x_{1,0} \\ x_{1,1} \\ \vdots \\ x_{1,k} \\ -- \\ x_{2,0} \\ x_{2,1} \\ \vdots \\ x_{2,k} \end{bmatrix} = \begin{bmatrix} y_{1,0} \\ y_{1,1} \\ \vdots \\ y_{1,k} \\ -- \\ y_{2,0} \\ y_{2,1} \\ \vdots \\ y_{2,k} \end{bmatrix} = \mathcal{P}\underline{Y}_k$$

where $\mathcal{P}$ is a permutation matrix. Hence, by reorganizing the vectorial signal inside the vectors $\underline{X}_k$ and $\underline{Y}_k$, the transformation is again lower triangular and unit diagonal.

- Let us note that Maison and Vandendorpe's approach can also be described by the following transformation

$$L\mathcal{P}\underline{X}_k = \begin{bmatrix} 1 & & & & 0 & & & \\ \star & \ddots & \mathbf{0} & & \star & \ddots & \mathbf{0} & \\ \vdots & \ddots & \ddots & & \vdots & \ddots & \ddots & \\ \star & \cdots & \star & 1 & \star & \cdots & \star & 0 \\ \star & & & & 1 & & & \\ \vdots & \ddots & \mathbf{0} & & \star & \ddots & \mathbf{0} & \\ \vdots & \ddots & \ddots & & \vdots & & \ddots & \\ \star & \cdots & \cdots & \star & \star & \cdots & \star & 1 \end{bmatrix} \begin{bmatrix} x_{1,0} \\ x_{1,1} \\ \vdots \\ x_{1,k} \\ -- \\ x_{2,0} \\ x_{2,1} \\ \vdots \\ x_{2,k} \end{bmatrix} = \begin{bmatrix} y_{1,0} \\ y_{1,1} \\ \vdots \\ y_{1,k} \\ -- \\ y_{2,0} \\ y_{2,1} \\ \vdots \\ y_{2,k} \end{bmatrix} = \mathcal{P}\underline{Y}_k.$$

The degrees of freedom corresponding to the triangular block (1,2) in Wong's approach have been transfered to the upper triangular block (2,1) in Maison and Vandendorpe's approach.

To precise this, let us consider a first causal transform $\underline{Y}_1 = L_1 \underline{X}$ with $R_{\underline{Y}_1\underline{Y}_1} = L_1 R_{\underline{X}_1\underline{X}_1} L_1^T = D_1$. Consider now another causal transformation $\mathcal{P}\underline{Y}_2 = L_2\,\mathcal{P}\,\underline{X}$ or $\underline{Y}_2 = \mathcal{P}^T\,L_2\,\mathcal{P}\,\underline{X}$ with $R_{\underline{Y}_2\underline{Y}_2} = (\mathcal{P}^T\,L_2\,\mathcal{P})\,R_{\underline{X}\underline{X}}\,(\mathcal{P}^T\,L_2\,\mathcal{P})^T = D_2$. Then

$$\det(D_2) = \det(R_{\underline{X}\underline{X}}) = \det(D_1) \tag{5.14}$$

The product of the variances of the subband signal is constant, no matter which causal transform we use, and as in chapter 2, the coding gain $G_{TC}^{(0)}$ is indeed invariant by permutation. Each permutation leads to another causal decorrelation of the components of one vector. For a stationary vectorial signal $\underline{x}$, this means that there exists more that one way to decorrelate the scalar signals which compose this signal. The examples of Wong, and Maison and Vandendorpe present in fact (for $M = 2$) two extreme cases of an infinity of variants, which are parametrized by the degree of (non) causality (in the classical sense) of the interband predictor(s).

5.3.4 Influence of the Noise Feedback

Let us now compare the approaches of Wong, and Maison and Vandendorpe in the presence of quantization. Prediction should be based on quantized data, which from section 5.2.2, perturbs the coding gain[5]. The expression (5.11) shows that in order to maximize the gain $G_L^{(1)}$, one should maximize the sum of the inverses of the optimal prediction error variances, $\sum_{i=1}^{M} \frac{1}{\sigma_{y_i}^2}$. Consider the case $M = 2$: let us assume, without loss of generality, that the variances $\sigma_{x_i}^2$ composing the vectorial signal are placed in decreasing order. In this case, one should minimize $\sigma_{y_2}^2$. This variance will be minimized if the largest number of

[5] In order to simulate a realistic source coding framework, one numerical result based on quantized data was presented in [86]; Wong [87] explicitly made the assumption that the power spectrum densities of the input of the crossband predictor and that of the corresponding unquantized signal are equal.

samples are used to predict $x_{2,k}$. The triangular approach of Wong should therefore be the best one, since it will lead to a smaller variance for $\sigma^2_{y_2}$. This difference between the two transformations appears only when the prediction is based on a quantized signal, but this is the way in which such decorrelating transforms will be implemented[6]. The following section deals with the general N case.

5.4 Optimal Ordering of the Subsignals for Closed Loop Triangular MIMO Prediction

Very few results, except from [88, 89], seem to concern an optimal ordering w.r.t. to the order of the decorrelation in prediction. These problems were presented in the framework of lossy [89] and lossless [88] coding of multispectral images; they are however different from those of investigated in our work. In [88, 89], one signal, or *band* (images of the same spectral band) is chosen as the best predictor (*anchor* band) for the other signals. This choice of a single anchor band is due to the constraints on the algorithmic complexity, which must be kept low in order to facilitate on-board implementation of the coding sheme in the spacecraft. In a second version of the algorithms proposed in these works, each previously compressed/decompressed band may be chosen as a possible predictor for the bands remaining to be coded, which poses the problem of an optimal ordering; computationnaly efficient solutions are then found using graphs theory. In the present work now, each remaining signal may be coded by *all* the previously coded/decoded signals.
Comparing $G_L^{(1)}$ in (5.11) with the infinite resolution case (5.9), the different variances produced by the different decorrelation approaches induce now different sums. Hence, the coding gain $G_{TC}^{(1)}$ depends on a carefull choice of the decorrelation procedure. In the case $M = 2$, maximizing the coding gain entails making the variances as different as possible. Thus, the subsignal of greater variance should be processed first, and all the degrees of freedom of the interband decorrelator should be used to decrease the variance of the subsignal of lower variance. The triangular MIMO predictor is in this case superior to the classical MIMO predictor, since W_{12} defined above is the most efficient interband predictor. Now for $M > 2$, the following theorem holds.

Theorem:**Optimal ordering of the subsignals for triangular MIMO prediction**. *The optimal ordering of the subsignals in a stationary vectorial signal for maximizing the high-resolution coding gain $G_{TC}^{(1)}$ of vectorial DPCM with triangular MIMO prediction is obtained by processing the signals in order of decreasing variance.*

To show the theorem, consider a recursive argument. First of all, the theorem is clearly true for the case of two channels. Now consider $n-1$ channels that we have ordered in order of decreasing variance. When we add a nth channel, the question is in which position it should be put w.r.t. the other channels. Assume in a first scenario that we put the channel in a position such that all n channels are in order of decreasing

[6] Another improvement due to Wong's approach appears when the filters are forced to have a finite length, see section 5.5.

variance. Assume in a second scenario that we insert the nth channel at another position. Then we can evolve from the first to the second scenario by a sequence of permutations of two consecutive channels. In one such permutation operation, assume that the channels involved in the permutation are in positions i and $i+1$. Then the channels $1, \dots, i-1$ are unaffected in the triangular MIMO prediction approach. The channels $i+2, \dots, n$ are also unaffected by the order in which channels i and $i+1$ are put since in any case they get orthogonalized w.r.t. the signals in those channels. So the only effect of the permutation between channels i and $i+1$ is on the prediction error variances of those channels i and $i+1$. In other words we are reduced to the two channel case, in which case we know that we should put the channels in order of decreasing variance. So, as we move from scenario one to scenario two by a succession of permutations of two consecutive channels, we decrease the coding gain in each permutation. Hence, the optimal ordering is in order of decreasing variance.

The (closed loop) triangular MIMO predictor can be seen as a generalization to the vectorial case of the classical (scalar) ADPCM coding technique, see figure 5.2.

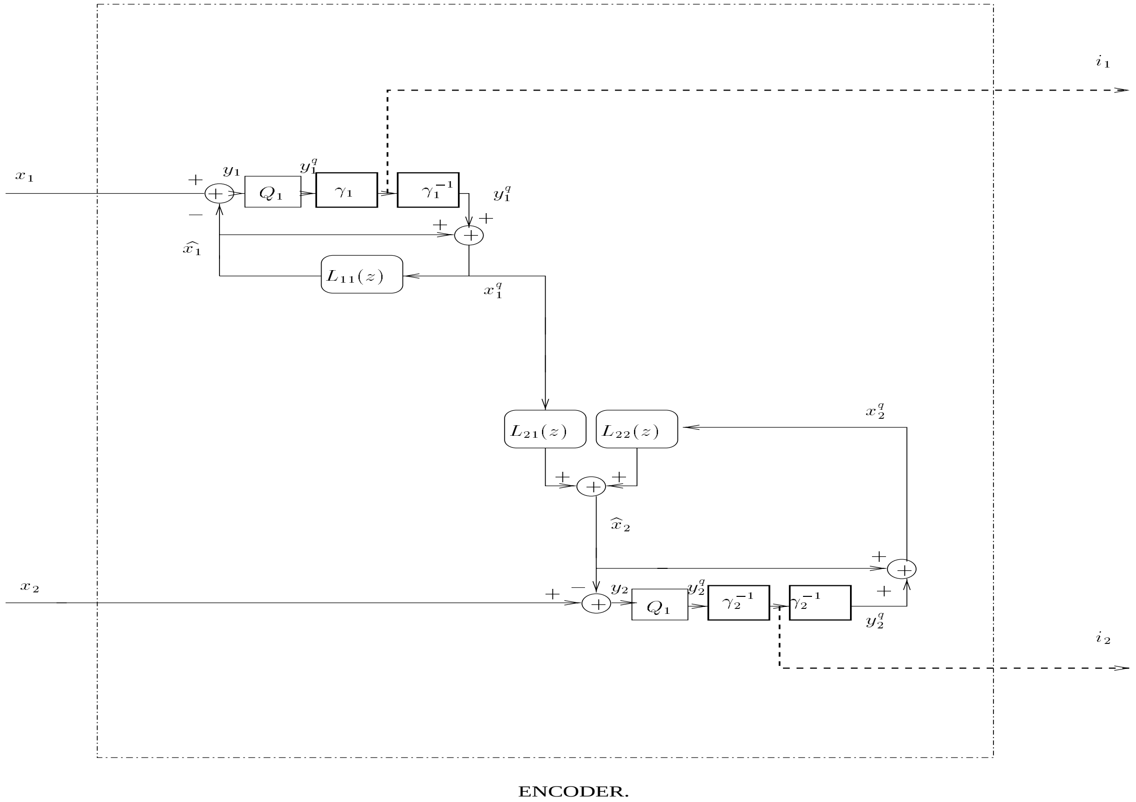


Figure 5.2: Encoder of the triangular MIMO predictor (“Vectorial DPCM”) for $M = 2$. The bitstreams i_1 and i_2 are transmitted to the decoder. Prediction of x_2 is non causal w.r.t x_1^q (through $L_{21}(z)$), and causal w.r.t. x_2^q (through $L_{22}(z)$).

In [55], this technique was therefore named ”VDPCM”, for vectorial DPCM. A possible confusion may however arise with *predictive quantization* [65], which uses vector prediction and vector quantization. We will therefore not retain this term in this thesis. The original VDPCM technique was first introduced by Cuperman and Gersho in [90], and was referred to as “vector DPCM”in [91], and as “differential vector quantization” (DVQ) in [92]. The coding technique introduced in the present work is different from these approaches because scalar, instead of vector quantizers are used to quantize the prediction residuals. Moreover, practical implementation of this scheme suggests closed loop implementation of the prediction, which in turn suggests a sequential (instead of block) procedure.

5.5 Optimal Triangular MIMO Prediction with Finite Prediction Orders

So far we have assumed that all filters involved are of infinite length. In the classical MIMO linear prediction, a finite number of prediction coefficients is typically used in a way that is a straightforward extension from the scalar case. Namely, the MIMO prediction order is limited to a finite order, resulting in a desired number of prediction coefficients (from the point of view of complexity or performance or both). In the triangular predictor case, it is more straightforward to assign a finite number of coefficients in an optimal fashion. The diagonal terms in the MIMO prediction filter correspond to classical scalar predictors, so the number of assigned coefficients will simply determine the prediction order as usual. However, for the non-causal off-diagonal terms, the filters are Wiener filters of unconstrained structure, except that we wish to use a finite number of taps. The problem then becomes the optimal positioning of those taps. In what follows, we shall assume that the diagonal scalar predictors are of sufficient order for the whitened versions of the signals to be considered as effectively white. In that case, the design of the off-diagonal terms in a row of the MIMO prediction filter corresponds to an issue of estimating a signal x on the basis of uncorrelated variables y_i. Due to the uncorrelatedness of the y_i, the estimation in terms of the y_i decouples and the contribution of each y_i can be considered separately. In particular, the variance of the estimation error becomes

$$r_{\tilde{x}\tilde{x}} = r_{xx} - \sum_i \frac{(r_{xy_i})^2}{r_{y_iy_i}} \tag{5.15}$$

where r_{xy} is the correlation. So, those variables y_i should be used for which the ratio $\frac{(r_{xy_i})^2}{r_{y_iy_i}}$ is the largest. Within a subset of the y_i that are samples of a certain whitened signal, $r_{y_iy_i}$ is independent of i due to stationarity and hence it suffices to use those samples y_i for which $|r_{xy_i}|$ is largest. The optimal positioning of a finite number of taps in the off-diagonal filters is therefore fairly straightforward.

5.6 MIMO Prediction of Audio Signals in the Frequency Domain

The several channels of a multichannel audio data stream may be stronlgy dependent, especially for recordings of an audio scene by multpiple microphones. For these audio signals, it has been experimentaly proved that coding schemes which remove the interchannel redundancy can considerably increase the coding efficiency [93]. This redundancy may be removed either in the temporal or in the frequency domain. In the latter case, the transform coefficients may be obtained by means of DFT, DCT, etc, applied to blocks of N samples of the scalar signals $x_{i,k:k+N-1}$. In this case, the transform coefficients belonging to the same channel are (almost) uncorrelated[7]. Thus, the interchannel redundancy removal can be performed by means of a decorrelating transfom (such as KLT or LDU) at each frequency band. Such an approach was proposed in [93, 95], where the frequency decorrelation was performed via KLTs, and added as a post processing stage in the core of a perceptual MPEG audio codec. This approach is attractive because it avoids the problem of finding the delay to which the channel are the most correlated[8]. For this reason, better compression was achieved by decorrelating the channels in the frequency than in the temporal domain in [95].

We will not pursue further this approach here; we should however note that, in the particular case of perceptual audio coders, the interchannel decorrelation performed in the temporal domain by means of Wiener filters should be carefully designed. The bitrate reduction caused by the decorrelation at low frequencies may be compensated by an excess bitrate due to the noise introduced by the predictors at higher frequencies; this may make the components at these frequencies more greedy in bits for the corresponding quantization noise to be maintained under the perceptual masking threshold [97].

5.7 Applications to Wideband Speech Coding

In the third generation mobile networks, the encoded signal band in wideband speech coders is 7kHz instead of the usual 3.4kHz. One way to construct such a coder is to filter and split the input signal into two subbands, which allows one to use an existing narrowband coder for the lowest subband ($x_{1,k}$, see figure 5.3.

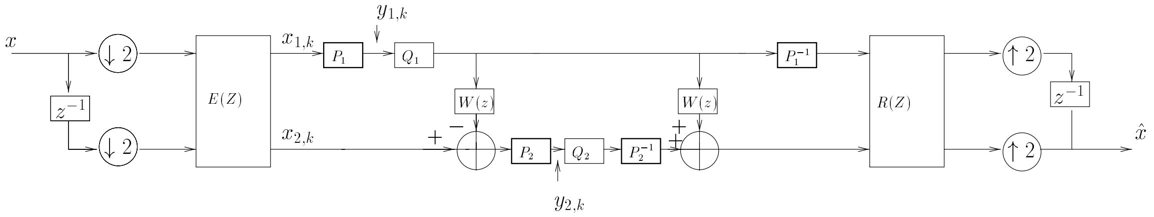


Figure 5.3: Triangular MIMO prediction applied to WideBand Speech Coding.

[7] This is only asymptotically true in the framelength N; see [94] for bounds on the coding gain of the DFT.

[8] Negligible instantaneous correlation exist in stereo signals, even in those generated by the recording of the same sonore source [96].

MIMO prediction may be applied in such a scheme to decorrelate the subband components $x_{1,k}$ and $x_{2,k}$. In the case of an optimal bit assignment, Wong's strategy described above should be applied: since the higher subband has on the average a lower variance than the lower subband, this approach should be the best decorrelating predictive transform. Note that, despite the non causality in the classical sense of this approach, it is well suited for frame based speech coding[9], which allows a certain degree of non causality. Actually, one can code one frame of signal in the lower subband and then code one frame in the higher subband.

Another special case is when the bit assignment is fixed, and when all the bits are used to code the lower subband. In this case, the quantization noises introduced by the quantization of the signals $y_{1,k}$ and $y_{2,k}$ are $\sigma^2_{\tilde{y}_1} = c2^{-2R}\sigma^2_{y_1} = \alpha\sigma^2_{y_1}$, with $\alpha \ll 1$, and $\sigma^2_{\tilde{y}_2} = \sigma^2_{y_2}$. The coding gain is

$$G_L = \frac{E\|\tilde{\underline{x}_k}\|^2_I}{\alpha\sigma^2_{y_1} + \sigma^2_{y_2}} \tag{5.16}$$

In this case again, the term $\alpha\sigma^2_{y_1}$ being small compared to $\sigma^2_{y_2}$, one has to minimize $\sigma^2_{y_2}$, and Wong's approach is more efficient. Informal listening tests we performed (using several GSM AMR narrowband codecs) have confirmed the perceptual gain over narrowband coding, introduced by the interband prediction. The prediction of the higher subband is done on the basis of the decoded version of the lower subband. The length of the frames was 200 samples in the subsampled domain, and the interband Wiener filter comprised of 21 taps (symmetric around the 0 lag). The filters were either the 32-taps QMF of Jain et al. [98], or the 32-taps CQF of Smith et al. [99]. The encoded lower subband is transmitted ($R_1 = 2R, R_2 = 0$) along with the coefficients of the crossband predictor (for which we assumed perfect quantization in these results). The decoder produces, by way of higher subband, only its predicted version on the basis of the decoded lower subband. The improvement in perceptual quality is nevertheless significant. Some overhead is required in transmitting the prediction filter $W_{21}(z)$, since backward adaptation is indeed made impossible in this case (there is no genuine highpass subband to predict).

This technique is closely connected to the problem of *bandwidth expansion*, for which the goal is the generation at the decoder of an acceptable upper subband subject to the constraint that no rate should be dedicated to its coding. The only available information about the high frequency band is therefore the quantized lower subband -assuming that significant statistical dependencies (or mutual information) exist between the two bands. In [100], a lower bound on the mean log spectral distortion (mLSD) of the spectral envelope in the missing frequency band as achievable by any memoryless bandwidth expansion algorithm is presented. The mLSD is first related to the mutual information shared between sets of parameters (the more the mutual information the less the mLSD). This information is then estimated for long term speech sequences, and for usual coding paramaters (LPÇ correlations coefficients,...). The minimal corresponding mLSD is evaluated at roughly 3 dB in the missing frequency band[10]. This paper provides also a detailed list of references about this interesting problem, including non memoryless techniques.

[9] such as that considered here, which uses GSM-AMR NB codecs

[10] As a rule of thumb, a “sufficient” WB speech quality corresponds usually to 1 dB on the average of the *total* band

5.8 Conclusions

For vectorial sources with memory, we showed in this chapter that the optimal causal decorrelating scheme can be described by means of a prediction matrix whose entries are optimal prediction filters. The diagonal filters are scalar intrasignal prediction filters. The off-diagonal predictors are Wiener filters performing the intersignal decorrelation. This decorrelating procedure led to the notion of “generalized MIMO prediction”, in which a certain degree of non causality may be allowed for the off-diagonal prediction filters. In the case of non causal intersignal filters, the optimal MIMO predictor is still lower triangular, and hence “causal”, in a wider sense. The notion of causality was generalized in the sense that causality between channels becomes processing the channels in a certain order. Some signals may be coded using the coded/decoded versions of the “previous” signals. We showed that two previously introduced transformations, in the context of subband coding, appear as special cases of the generalized MIMO prediction. As the previously described causal LDU transform, realistic coding implementations of the latter two approaches should involve closed loop structures for the prediction. We showed that though these approaches are equivalent in the limit of high rates, triangular MIMO prediction may be more efficient than its classical counterpart. This triangular predictor appears as an extention of the classical scalar (A)DPCM to the vector case. In this case, we showed that the optimal ordering of the scalar signals (w.r.t. the coding performance at high rate) corresponds to the case where they get decorrelated by order of decreasing variances. In the case where FIR filters are used to perform the prediction, the triangular predictor was shown to benefit from a simple optimal positioning of the taps for the off-diagonal filters. Finally, we presented some applications of these results to wideband speech coding.

5.A Derivation of (5.8)

The determinant of the lower triangular and unit diagonal prediction matrix $L(f)$ may be written as

$$\det L(f) = \prod_{i=1}^{M} L_{ii}(f), \tag{5.17}$$

from which we obtain

$$\begin{aligned} \textstyle\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln \det[L(f)]\, df &= \sum_{i=1}^{M} \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln L_{ii}(f) df \\ &= \sum_{i=1}^{M} Q_i(f). \end{aligned} \tag{5.18}$$

It is now shown that any of the M previous integrals $Q_i(f)$ is zero. The idea of the proof is to show that these sums do not depend on the coefficients of the prediction filters, and in particular, they may be set to 0 without affecting the result.

Since $L_{ii}(z) = \sum_{k=0}^{\infty} L_{ii,k} z^{-k}$ are prediction filters, they have causal and stable inverses $B_{ii} = \sum_{k=0}^{\infty} B_{ii,k} z^{-k}$. Thus we have

$$\begin{aligned} \tfrac{\partial Q_i(z)}{\partial L_{ii,k}} &= \textstyle\oint \frac{\partial L_{ii}(z)/\partial L_{ii,k}}{L_{ii}(z)} \frac{dz}{z} \\ &= \textstyle\oint z^{-k} B_{ii}(z) \frac{dz}{z} \\ &= \oint z^{-k} \sum_{j=0}^{\infty} B_{ii,j} z^{-j} \frac{dz}{z} \\ &= \sum_{j=0}^{\infty} B_{ii,j} \oint z^{-(k+j)} \frac{dz}{z} \\ &= \sum_{j=0}^{\infty} B_{ii,j} \delta_{0,j+k} \\ &= B_{ii,0} \\ &= 1. \end{aligned} \tag{5.19}$$

Hence, $Q_i(z)$ does not depend on the strictly causal coefficients $L_{ii,k}$, and is therefore equal to that obtained with $L_{ii}(z) = 1$, which is zero.

5.B Derivation of (5.12)

Similarly to section 2.C, we consider the optimal decorrelation of $R_{\underline{X}_k \underline{X}_k} + \sigma_q^2 I_{kM}$. Then we have

$$\begin{aligned} \lim_{k\to\infty} \left(\det\left[\text{diag}\left\{L'(R_{\underline{X}_k \underline{X}_k} + \sigma_q^2\, I_{kM})L^{'T}\right\}\right]\right)^{\frac{1}{k}} &= \prod_{i=1}^{M} (\sigma_{y_i}^2 + \Delta\sigma_{y_i}^2 + \sigma_q^2), \\ &= \prod_{i=1}^{M} \sigma_{y_i}^2 \left(1 + \frac{\Delta\sigma_{y_i}^2 + \sigma_q^2}{\sigma_{y_i}^2}\right) \end{aligned} \tag{5.20}$$

where $\Delta\sigma^2_{y_i}$ correspond, as in 2.35, to the increase in prediction due to the noise feedback (w.r.t. to the optimal prediction error variances $\sigma^2_{y_i}$), and I_{kM} is the $kM \times kM$ Identity matrix. Since L' totally decorrelates $R_{\underline{X}^q_k \underline{X}^q_k}$, we have

$$\begin{aligned} \prod_{i=1}^{M} \sigma^2_{y_i} \left(1 + \frac{\Delta\sigma^2_{y_i} + \sigma^2_q}{\sigma^2_{y_i}}\right) &= e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\left[\det(S_{\underline{xx}}(f) + S_{\tilde{y}\tilde{y}}(f))\right] df} \\ &\approx e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\left[\det(S_{\underline{xx}}(f))\right] df} \left(1 + \int_{\frac{1}{2}}^{\frac{1}{2}} \operatorname{tr}\left(S^{-1}_{\underline{xx}}(f) S_{\underline{\tilde{y}\tilde{y}}}(f)\right) df\right) \\ &\approx \left(\prod_{i=1}^{M} \sigma^2_{y_i}\right) \left(1 + \sum_{i=1}^{M} \frac{\Delta\sigma^2_{y_i} + \sigma^2_q}{\sigma^2_{y_i}}\right) \end{aligned} \tag{5.21}$$

where tr denotes the trace operator. Now, the required quantity for the coding gain is

$$\begin{aligned} \prod_{i=1}^{M} \sigma^2_i &= \prod_{i=1}^{M} (\sigma^2_{y_i} + \Delta\sigma^2_{y_i}) \\ &\approx \left(\prod_{i=1}^{M} \sigma^2_{y_i}\right) \left(1 + \sum_{i=1}^{M} \frac{\Delta\sigma^2_{y_i}}{\sigma^2_{y_i}}\right), \end{aligned} \tag{5.22}$$

which from (5.21) may be written as

$$\left(\prod_{i=1}^{M} \sigma^2_{y_i}\right) \left(1 + \sum_{i=1}^{M} \frac{\Delta\sigma^2_{y_i}}{\sigma^2_{y_i}}\right) \approx e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\left[\det(S_{\underline{xx}}(f))\right] df} \left[1 + \int_{\frac{1}{2}}^{\frac{1}{2}} \operatorname{tr}\left(S^{-1}_{\underline{xx}}(f) S_{\underline{\tilde{y}\tilde{y}}}(f)\right) df - \sum_{i=1}^{M} \frac{\sigma^2_q}{\sigma^2_{y_i}}\right]. \tag{5.23}$$

Setting $S_{\underline{\tilde{y}\tilde{y}}}(f) = \sigma^2_q I$, we obtain for the coding gain $G^{(1)}_{TC}$

$$\begin{aligned} G^{(1)}_{TC} &= \left(\frac{\prod_{i=1}^{M} \sigma^2_{x_i}}{\prod_{i=1}^{M} \sigma^2_i}\right)^{\frac{1}{M}} \approx G^{(0)}_{TC} \left[1 + \sigma^2_q \int_{\frac{1}{2}}^{\frac{1}{2}} tr\left(S^{-1}_{\underline{xx}}(f)\right) df - \sum_{i=1}^{M} \frac{1}{\sigma^2_{y_i}}\right]^{-\frac{1}{M}} \\ &\approx G^{(0)}_{TC} \left[1 + \frac{\sigma^2_q}{M} \left(-\int_{-\frac{1}{2}}^{\frac{1}{2}} \operatorname{tr}\left(S^{-1}_{\underline{xx}}(f)\right) df + \sum_{i=1}^{M} \frac{1}{\sigma^2_{y_i}}\right)\right], \end{aligned} \tag{5.24}$$

which is the desired expression.

Brief History of this work: Analysis-by-Synthesis Structures

The results of the presented work find their origin in the french RNRT project *COBASCA*[11], which aimed of providing source and (joint source-) channel coding algorithms for wideband audio signals ($[50Hz - 7kHz]$) in the framework of UMTS. Our personal contribution to this project concerned mainly source coding, for which we followed two axis of research. The first axis lies beyond the scope of the present framework, and will only be briefly summarized. The description of the second axis may however be relevant for the reader interested in the topics of this thesis; it is shown how existing source coding techniques, industrial constraints and scientific objectives together led to the causal coding framework described along these pages.

Joint Optimization of Formant and Pitch Predictors

Low bit rate speech coding makes a pervasive use of linear prediction. The GSM AMR codecs standardized by ETSI for narrowband speech coding are based on CELP algorithms. As most source codecs, decorrelation is performed (by means of linear prediction) before entropy coding. Due to the particular structure of speech waveforms, the prediction is comprised of two stages: a short term predictor (STP) removes short terms correlations (*formants* due to the vocal tract), and a long term predictor (LTP) deals with more distant correlations due to the excitation of the vocal cords (*pitch*). Though the coding algorithms of CELP coders may be very elaborated, they *separately* estimate these predictors. The depicted correlations are however not independent, and a sequential approach for linear prediction is not optimal. We proposed therefore a method allowing one to jointly estimating STP and LTP, using an iterative algorithm. In order to establish the efficiency of this joint optimization, an analysis-by-synthesis -like criterion was proposed. An estimate of the original signal is computed filtering the excitation (white noise) by means of the jointly estimated predictors; this signal is then compared to the original which is a synthetic stationary signal whose optimal STP and LTP are perfectly known. The results show that a joint optimization clearly improves the decor-

[11] *CO*dage en *B*ande élargie avec partage *A*daptatif du débit entre *S*ource et *CA*nal pour réseaux cellulaires de deuxième et troisième générations (UMTS), http://www.telecom.gouv.fr/rnrt/pcobasca.html.

relation efficiency. Moreover, both the complexity of the proposed method, and the number of iterations required by the iterative algorithm to converge to the optimum estimates are fairly low. Details about this work may be found in [101, 102].

Analysis-by-Synthesis Coding of Wideband Speech

The second axis regarding wideband coding of speech was based on the idea of using an existing (standardized) narrowband coder; this approach was attractive because it would indeed greatly simplify the optimization work concerning the coding of the lower (and most important) subband $[50Hz - 4kHz]$. The remaining problem was that of designing a convenient filter bank. Besides traditional constraints such as delay and passband selectivity, preliminary results showed that the quality of the codecs of the GSM AMR-NB rapidly decreases beyond $3.4kHz$. Moreover, this frequency may be within the frequency area where symmetric two channel filterbanks overlap for $8kHz$ original signals (aliasing may be accounted for by the relation of analysis to synthesis filters, but reappears because of the quantization). Finally, the human hear is particularly sensitive in this frequency region. These facts suggested the use of an analysis-by-synthesis technique, which had proven usefull results by the past in speech coding[12]. An attractive structure was therefore that of figure 5.4, based on the Laplacian Pyramid [103] (in the figures of this chapter, Q_i may denote any codec). The input signal is wideband ($[50Hz - 7kHz]$). The filters H_0, G_1, H_1', G, and G_0' should now be optimized subject to the constraint of minimizing the variance of the reconstruction error $E(x - x_r)^2$. First, we need one branch to be a genuine lowpass subband, which fixes consequently H_0. Second, by writing explicitly the polyphase components [104] of the signals and those of the filters (*e.g.*, $X^e(z)$ and $X^o(z)$ denote respectively the even and odd components of the input signal, and similarly for the filters and the reconstructed signal), we may obtain more insight about the role plaid by the other filters. The polyphase analysis leads to[13]

$$\begin{aligned} X_r^e &= X^e[G_0^{'e}H_0^e + G_1^eH_1^{'e} - G_1^eH_1^{'e}G^eH_0^e - G_1^eH_1^{'o}G^oH_0^o] \\ &\quad +z^{-1}X^o[G_0^{'e}H_0^o + G_1^eH_1^{'o} - G_1^eH_1^{'e}H_0^oG^o - G_1^eH_1^{'o}G^oH_0^e], \\ X_r^o &= zX^e[G_0^{'o}H_0^e + G_1^oH_1^{'e} - G_1^oH_1^{'e}G^eH_0^e - G_1^oH_1^{'o}G^oH_0^e] \\ &\quad +X^o[G_0^{'o}H_0^o + G_1^oH_1^{'o} - G_1^oH_1^{'e}H_0^oG^e - G_1^oH_1^{'o}G^oH_0^o]. \end{aligned} \tag{5.25}$$

Considering the analysis-by-synthesis branch $G\ H_1'$ in fig. 5.4 (b), ne may denote by F' the filter equivalent to the cascade

$$F' = G^eH_1^{'e} + G^oH_1^{'o}. \tag{5.26}$$

This relation expresses the influence of one branch uppon the other, see figure 5.5.

[12] CELP coders are *e.g.* well known analysis-by-synthesis coders: the excitation signal is selected among several candidates of a codebook by synthesizing the reconstructed speech (through the inverses of the LTP and STP), and by choosing the most representative w.r.t. to a weighted distortion measure.

[13] dependence in z is omitted for notation simplicity.

Comparing these representations with figure 5.4 (a), this figure shows in particular that the analysis-by-synthesis filterbank may be linked to a classical filterbank, where respectively $G'_0 - F'(z^2)G_1$ and $H'_1 - F'(z^2)H_0$ would be identified to respectively G_0 or H_1 QMF- or CQF- like filters. Focusing then on the role of the filter G, it should be designed in order to minimize the variance of the input of the quantizer Q_2, as shown by the substractive branch of Figure 5.4, (b). This branch allows one to optimize the whole structure through the optimization of the filter F' (figure 5.6), which corresponds to the cascade $G - H'_1$. Suppose we fix H_0, G'_0, H'_1 and G_1 as those of a classical filterbank (fig. 5.4, (a)). The remaining degrees of freedom are then the coefficients of the transfer function of G, which may be written as

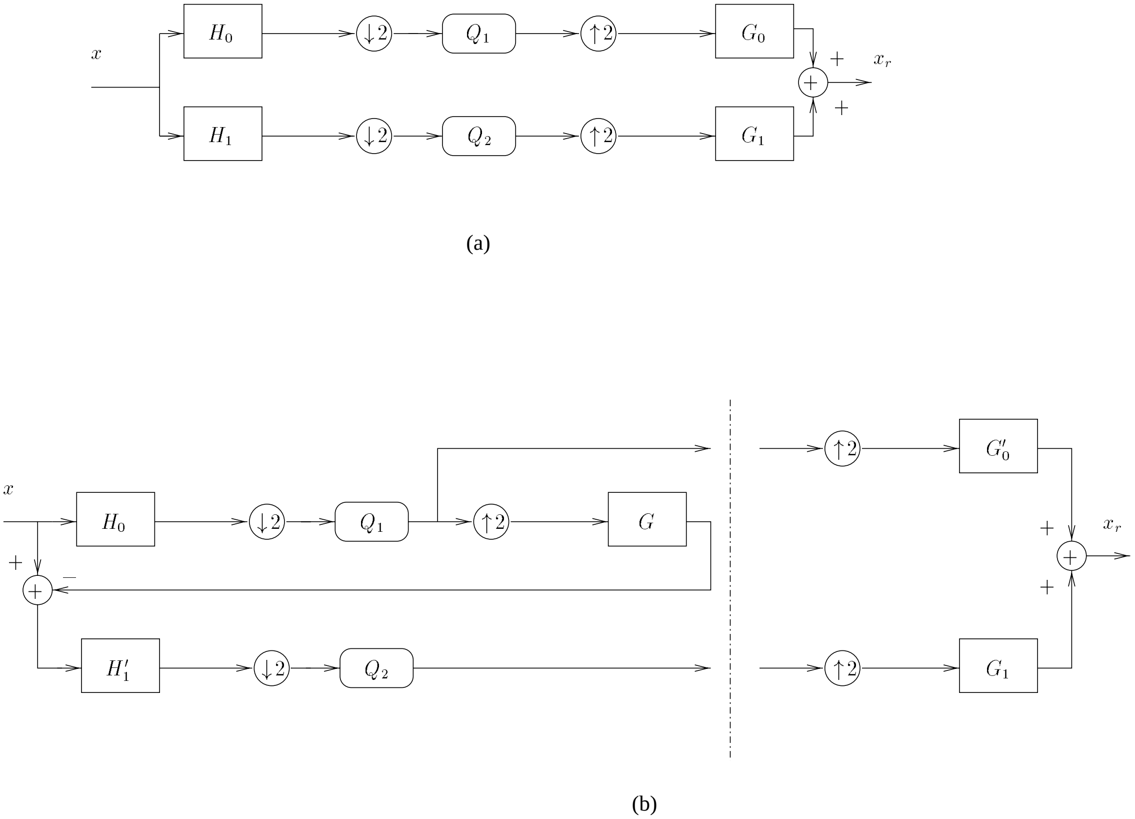
$$\begin{aligned} G^e &= G^e_0 F_{w_1} \\ G^o &= G^o_0 F_{w_2}. \end{aligned} \tag{5.27}$$

Using the expression (5.26) relating the components of F' to the analysis and synthesis filters, we obtain

$$F' = H^e_1 G^e_0 F_{w_1} + H^o_1 G^o_0 F_{w_2}, \tag{5.28}$$

where the components F_{w_1} and F_{w_2} of some filter F are the remaining degrees of freedom of the system. Thus, they correspond to adaptive Wiener filters aimed of modeling one subband signal on the basis of the other. This lead to the alternative representation of figure 5.7, where a crossband predictor should be designed to minimize the variance at the input of the second quantizer, which makes the subbands ideally decorrelated. The adaptive part of the whole structure is concentrated into this single filter; the other analysis and synthesis filters are those of a classical, and possibly separately optimized filterbank. The positioning of F w.r.t. to the quantizer Q_1 is therefore a consequence of the desired analysis-by-synthesis configuration. This is the structure described by Wong in his 1997's paper [87].

The decorrelation matrix resulting from this approach was expressed in 5.13. The link with the causal LDU transform was then straightforward, since a triangular matricial transform whose rows are optimal prediction filters (with increasing prediction orders) of the input signal diagonalizes the covariance matrix of these data; this renders the structure optimal in the classical high rate transform coding framework. The analysis-by-synthesis constraint led then to the closed loop implementation described in the first chapters of this thesis.



(a)

(b)

Figure 5.4: (a) Classical filterbank and (b) Laplacian pyramid-like structure applied to wideband coding of speech.

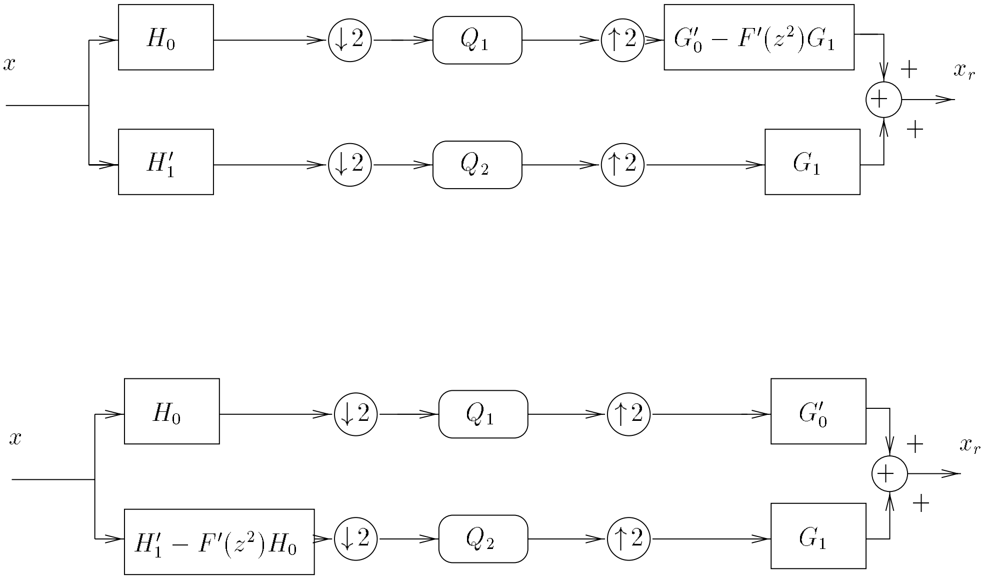


Figure 5.5: Equivalent representations of the analysis-by-synthesis filterbank.

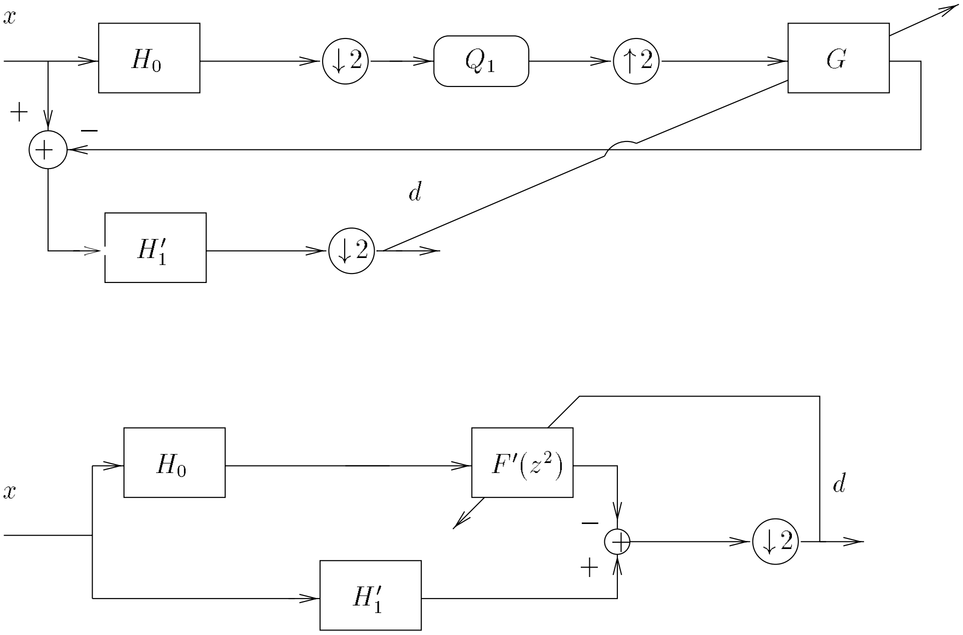


Figure 5.6: Equivalent optimizations of the analysis-by-synthesis filterbank w.r.t. the impulse response of G.

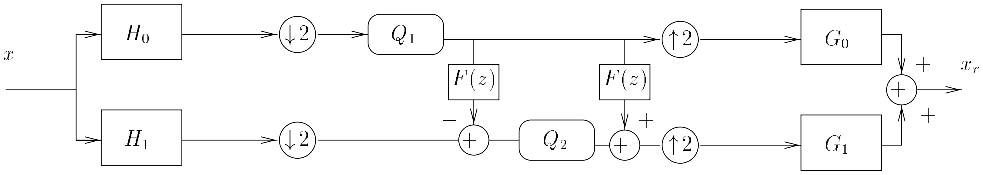


Figure 5.7: Representation of the analysis-by-synthesis filterbank as the Wong's structure.

Part II

Causal Lossless Coding

Overview of the Second Part

The second part of this thesis presents and analyzes lossless coding techniques based on the causal decorrelating approaches (LDU transform and generalized MIMO prediction) described in the first chapters. This overview is organized in four parts. The first one sets the stage; it presents the coding structures and the related problem which will be investigated. Basically, these structures involve integer-to-integer transforms, and multi-stage lossless coding. The second part presents a brief overview of state-of-the-art coding techniques and issues in lossless coding of multichannel audio, which is a natural field of application for the proposed coding procedures. The third part presents in more details the framework of multiresolution coding. Finally, the last part details in more depth the particular contents of each chapter.

Framework and Coding Structures Analyzed in this Part

Integer-to-Integer transforms

Lossless coding schemes may exist as stand-alone encoders, but they are also part of the core of lossy encoders, in order to improve its compression efficiency; this is the goal of the entropy coding techniques described in the introduction of this thesis. Let us consider now the coding scheme depicted in figure 5.8, which uses a decorrelating transformation T.

In a first step, a very high resolution (amplitude-continuous) vectorial source $\underline{x}^c$ is quantized using a lossy source codec, represented by the box Q (Q may represent the discretization realized by any lossy codec, *e.g.* independent uniform scalar quantizers, independent ADPCM or MPEG audio codecs...). Once the quantization has been performed, one is left with a discrete-valued vector source $\underline{x}$. The problem is then to transmit efficiently (w.r.t. the bitrate and the complexity) the vectors $\underline{x}_k$ to the decoder. An efficient entropy coding procedure is vector entropy coding; it is known to be asymptotically optimal w.r.t. to the block length, but requires to estimate the joint probability distributions of the vectors. Such a coding procedure is consequently very complex and not well suited to signals which present long term correlations[14] (such as high quality audio signals sampled at 44.1kHz). In this case, the set of streams $\{x_i\}$ obtained from the

[14] For vector sources with memory, the problem is even more acute since joint probability distributions of vectors *of vectors* should be estimated.

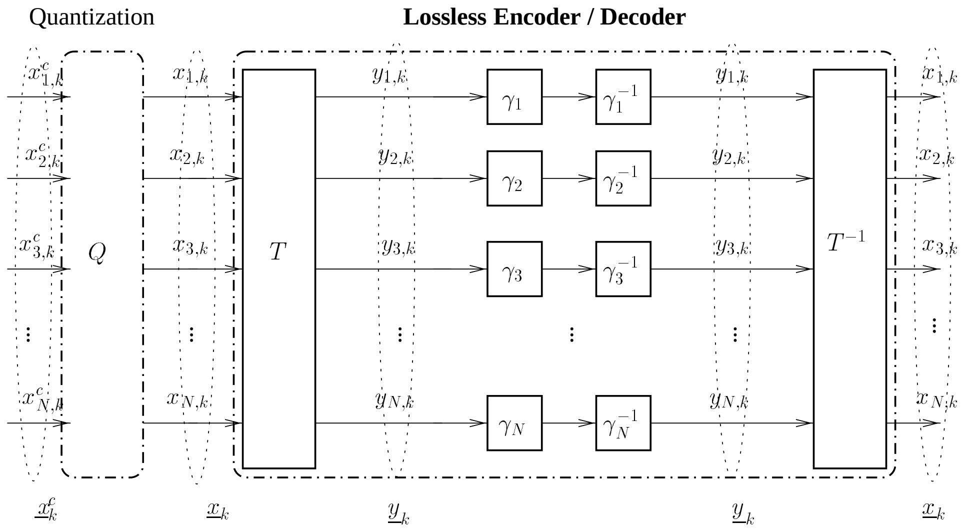


Figure 5.8: Transform based lossless coding scheme embedded in a lossy codec.

lossy encoder is preferably entropy coded using N scalar entropy coders γ_i[15]. However, the scalar sources x_i are generally neither memoryless, nor independent, which makes the single-letter entropy coding suboptimal. One may therefore apply after the quantization stage a lossless transformation T in order to reduce the intra- and inter- signal correlations, and thereby, the bitrate. Indeed, the signal should not incur further degradation: T must be invertible. Both the inputs and the outputs of T are discrete valued; therefore, T is called an integer-to-integer transform.

Summarizing this framework, the integer-to-integer approach divides the coding procedure into two steps: a transform T is firstly applied to each block in the aim of decorrelation; the transform components are secondly scalar entropy coded, which keeps the complexity reasonnably low. The vectorial signal $\underline{x}$ gives rise to N transform signals y_i from which the decoder is able to losslessly recover the original signal. This approach will be referred to as "one-shot", or "single-stage" lossless coding.

For a given transform T, and a given source $\underline{x}$ we will consider two scenarios: scenario 1, where T is used, and scenario 2, where it is not. In both cases, the structure will use N scalar entropy coders γ_i as in figure 5.8. We will then investigate the following questions. Firstly, what is the maximum achievable bitrate reduction over scenario 2? Secondly, what is the actual bitrate reduction operated by using the transform T? In chapter 6, T will be based on two decorrelating transforms: the KLT and the LDU. In chapter 8, T will be based on the MIMO decorrelating approaches discussed in chapter 5, which account for both intra- and inter-signal correlations.

[15] For example, popular codes in audio include Huffman and Golomb-Rice codes.

Multi-Stage Lossless Coding

Besides this "one-shot" compression approach, a different lossless coding procedure consists in lossy coding the source $\underline{x}$ in a first step, producing thereby a first streams of N "low resolution" signals y_i^q. In a second step, the error signal $\underline{e}$ is separately encoded, which results in the two-stage structure of figure 5.9.

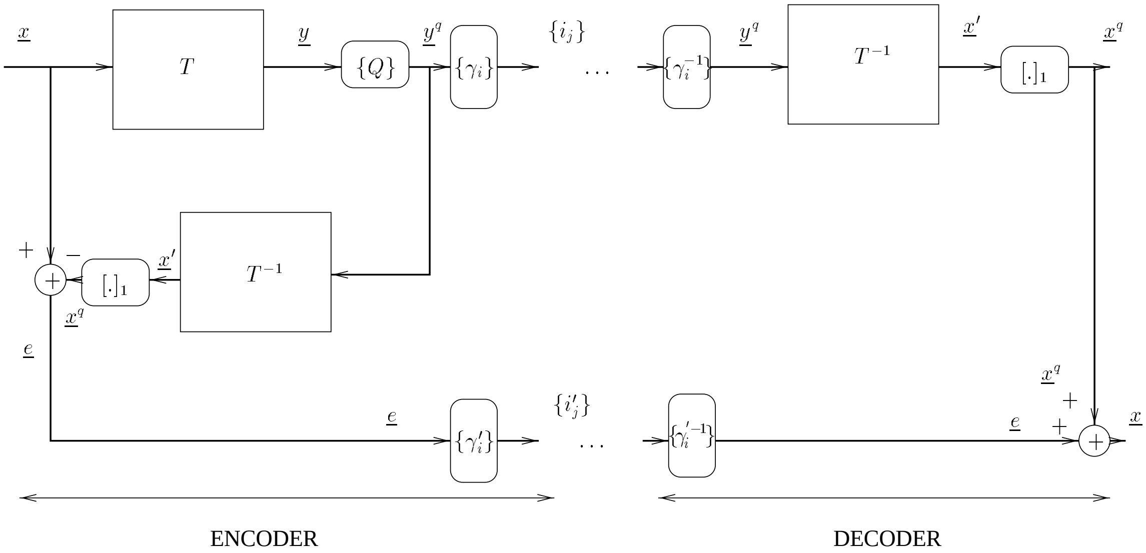


Figure 5.9: Classical two-stage lossless transform coding. $\{Q\}$ denotes uniform scalar quantizers, $\{\gamma_i\}$ and $\{\gamma_i'\}$ scalar entropy coders, and $[.]_1$ rounding operators.

The advantage of this scheme (*e.g.* in the case of variable transmission bandwidth, or internet browsing) is that an approximative version of the signal of interest can be quickly obtained, independently of the error signals. The original signal can eventually be recovered by adding the error signals. Depending on the stepsizes of $\{Q\}$, the rate dedicated to code the low resolution version $\underline{x}^q$ of $\underline{x}$ can be regulated; this permits for this signal lower rates than in a single-stage lossless coder, at the cost of introducing some distortion. This coding scheme is widely used in lossless coding of audio signals, see e.g. [21, 24], and of images [105, 106]. A comparison of the compression efficiency of standard orthogonal tranforms to that of the causal one appears therefore interesting. In particular, it is interesting to know wether using a two-stage lossless transform coding scheme is suboptimal w.r.t. to the single-stage approach explained above. These questions are addressed in chapter 7 for the two-stage approaches based on LDU and orthogonal transforms, and in chapter 8 for two- and M-stages structures based on MIMO predictors. We present now an overview of lossless multichannel audio, for which the considered coding schemes may be useful. Then, the concept of coding a source by means of multiple resolution levels will be described more precisely. It has a long history in source coding.

Lossless Multichannel Audio Coding

In the last decades, little attention has been paid to lossless audio coding, mainly because it provides lower compression ratios than lossy coding. Many modern applications suggest however the use of a powerfull lossless audio coding technique. In applications where coding is not subject to stringent bitrate constraints, as for Digital Versatile Disks (DVD), lossless coding obviously appears as the best technique. Some applications of very high fidelity music distribution over the internet could also provide lossy compressed audio clips in a first step (allowing the music lover to browse and select the desired clip in a reasonnable time), and then provide losslessly compressed audio signals in a second step. Such systems are called *scalable*[16] systems, and will be the topics of chapters 7 and 8. For archiving and mixing applications, lossless compression avoids signal degradation when successively encoded/decoded with lossy encoders [107]. It can also be observed that an increasing number of companies now provide products for lossless audio compression [108]. A complementary survey to that of [107], reviewing free competitive lossless codecs, can be found in [109]. In the particular case of MPEG-4, MPEG members are now discussing issues in considering lossless audio coding as an extention to the MPEG-4 standard [110].

An important issue for which lossless audio coding schemes should account is the multichannel aspect of recent audio technologies. Starting from the monophonic and stereophonic technologies, new systems (mainly due to the film industry and home entertainments systems) such as quadraphonic, 5.1 and 10.2 channels are now available. An efficient coding procedure aimed of storage, or transmission of these signals should benefit (sub)channel correlations.

Multichannel audio sources can be roughly classified into three categories : signals used for broadcasting, where the channels can be totally different one from another (e.g. different audio programs in each channel, or the same program in different languages), film soundtracks (typically the format of 5.1 channels) which present a high correlation between certain channels, and finally multichannel audio sources resulting from a recording of the same scene by multiple microphones (in this case, there is indeed a great advantage to be taken from the structure of the multichannel audio signal) [93].

In most state of the art lossless (and lossy) audio codecs however, interchannels correlations are not fully exploited; these systems often only compute sums and differences. This assertion should be contrasted by the recent works in [93] and [21], where KLT and adaptive prediction are respectively used to remove inter- and intra-channel redundancies. The former was evocated in chapter 6; the latter will be described more extensively in chapter 8.

Besides purely lossless systems, interesting alternatives are lossy/lossless coders. These systems either switch from lossy to lossless algorithms, or provide a lossy version of the signal first, and the complementary error signal in a later stage, resulting in multiresolution systems.

[16] The term *progressive* is more frequent in image coding.

Multiresolution Coding

The principle of multiresolution coding is that only an incremental increase in rate over the current transmitted rate results in an improvement in the source representation[17]. Depending on the available resources (transmission bandwidth, capacity storage), the scheme may be lossy or lossless, and provides an SNR scalability.

Let us consider coding methods designed to operate at fractional rates of an overall rate. The question whether this system is suboptimal in the rate-distortion sense w.r.t. the same system designed for the overall rate has been fist addressed from the rate-distortion theory viewpoint by Koshelev [112] who called it *divisibility*, Equitz and Cover [113] under the heading of *successive refinement information*. It is shown that successive refinement in the rate-distortion (r(D)) optimal sense is not always possible, and that a sufficient and necessary condition is that the individual encodings (or representations) be expressible as a Markov chain. More recently, this result was reinterpreted by Rimoldi [114] and extended from memoryless to more general sources in [50]. We will restrict the rate-distortion considerations of this second part by focusing on the operational multiresolution compression performance obtained by particular multiresolution (or multi-stage) coders only; these performance will be compared with the corresponding one-shot (or single-stage) lossless coders.

Progressive coding has become important in image and audio coding, since in a network environment, different users may have different access capabilities, such as different bandwidth, CPU power, etc, and may access the sources at different levels of quality. In such circumstances, a coder that can provide a coded sequence in a progressive way has an advantage. Progressive coding is also designated as scalability, multiresolution, layered or embedded coding, information divisibility, or successive approximation. Because it has become ubiquitous in practical coding systems, it is difficult to exhaustively present the several related techniques. Two basic approaches can however be distinguished: *spectral selection*, and *successive approximation* or *refinement*, which will be investigated in this work.

Spectral selection uses the signal representation obtained by means of a transform or a subband coder. Since for many signals (e.g. images or long term speech) most activity is concentrated in the low frequency area, an acceptable representation may be obtained by means of the corresponding (or, more generally, by the most significant) coefficients only[18]. This approach is for example used in MPEG audio codecs, where the significance of the transform coefficients[19] is computed w.r.t. to a *psychoacoustic mask*. This is aldo somewhat in the spirit of the AMR-WB codec where the high frequency band $[6.4 - 7]kHz$, which is not perceptually critically relevant, is discarded from the transmission [20] [116, 117]. As for images, many

[17] This contrasts with the Multiple Description framework, where the division of the overall rate is aimed of ensuring an acceptable quality in case of channel impairments [111].

[18] The *Significance map* locates for example the significative transform coefficients on a grid, and is transmitted as side information in the JPEG standard.

[19] In layers 1 and 2 of MPEG1, these coefficients are obtained by means of QMF; In layer 3, also called *MP3*, and MPEG2-AAC, by means of MDCT [115].

[20] A bandlimited white noise is instead spectrally shaped at the decoder, according to the formant structure of the lower frequencies.

approaches exist, including e.g. the prioritized DCT method [118], and lossless approaches based on Laplacian Pyramid [103] such as [119, 105, 80], on subband coding with QMF [120], antialiasing filters [121], or wavelets [122, 123].

Another powerfull progressive coding scheme is successive approximation. In contrast to spectral selection, which generates minimum distortion for the selected coefficients but discards all the other coefficients, successive refinement produces relatively constant distortion for all the coefficients. *Embedded coding systems* have the feature that bit rate reductions can be performed at any point along the communication network. They imply a block of bits within which is embedded a subblock, which is itself sufficient for producing a decoded signal of sufficient quality, although full quality is achieved only upon receiving the entire block. References about early systems may be found in [14]. PCM is for example a naturally embedded system (least significant bits are simply discarded first), but DPCM is not. In order to cope with possible degradations of the reconstructed signal, a tractable approach is to decrease the precision in the feedback loop: in this case, only the *core* bits (as opposed to *enhancement* bits) of the value of each quantized sample are used in the prediction. Other approaches are based on adaptively allocating the bits among the quantizer of the prediction residual and the quantizer for the reconstruction error [124, 125]. CCITT Recomandation G.727 describes embedded (A)DPCM algorithms using 5,4,3 and 2 core bits [126, 127]. A both bit rate and bandwidth scalable CELP coder is standardized in MPEG4 [128]. Besides, combination of (A)DPCM and spectral selection with Laplacian pyramid was studied in [129], and with filterbanks for lossy speech coding [130]. In images, examples of coders which use successive refinements (based on DWT) are the EZW (Embedded Zero Tree) algorithms of [131, 132]. Context information is used in successive refinement of image coding in [133, 134], and more recently in [135].

A two-stage lossless coder, including a previously standardized MPEG codec in the lossy stage was proposed in [136], and extended to multiple bit rates in [137]. Spectral selection and successive refinement may also be combined, as in [138]. A comparison between the performance of these various techniques can be found in [139, 140] and [141].

Proposed Analyses

The following topics will be investigated in this second part.

- Chapter 6 deals with single-stage transform coding. In the case where T of fig. 5.8 is based on decorrelating matrices such as KLT or LDU, the relation of $\underline{x}^c$ to $\underline{y}$ is similar to that obtained with transform coding, except that quantization and transform stages are *reversed*. The transformed signals must be discrete since they are further entropy coded. Therefore, integer-to-integer implementations of transforms traditionally used in the context of transform coding may be useful in such a scheme. This chapter will compare the compression performance of the KLT and the LDU in this framework. From a rate-distortion point of view, the question of whether integer-to-integer transforms are, as efficient as their continuous counterparts was addressed recently in [41]. Let us Assume that the quantization stage Q is comprised of uniform scalar quantizers with the same stepsize Δ. We will refer to the following coding schemes:

- (1) scalar quantization of the x_i^c followed by scalar entropy coders,
- (2) scalar quantization of the x_i^c followed by integer-to-integer decorrelating (single-stage lossless transform coding), transform and scalar entropy coders,
- (3) continuous decorrelating transform followed by scalar quantization and scalar entropy coders (transform coding),
- (4) quantization of the x_i^c followed by vector entropy coders.

The results of [41] show that, for Gaussian vectors, the performance of the schemes (2), (3), and (4) are equivalent in the limits of small stepsizes. This is equivalent to neglecting the integer-to-integer constraint on the transformation of scheme (2). The purpose of this chapter is to evaluate the bitrate reduction actually operated by scheme (2) w.r.t. scheme (1), when this constraint is accounted for. This bitrate reduction is defined as a *lossless coding gain*. We will show how the gains of schemes (3) and (4) represent an upper bound for that of(2) in terms lossless coding gain; this bound will be linked to the mutual information shared by the x_i. For a given quantization stage (fixed distortion level), the suboptimality of (2) will be expressed in terms of excess bitrate. This inherent suboptimality of integer-to-integer transforms will then be compared for the LDU and the KLT. The LDU will be shown to outperform the KLT in this case, because of its triangular structure. Finally, the adaptivity of the considered single-stage lossless transform coding systems will be investigated. This part is somewhat in the spirit of the analyses of chapter 5. We will consider systems whose integer-to-integer transforms are computed in a backward adaptive manner, by means of an estimate of the covariance matrix based on K decoded vectors. In this case, the lossless transforms converge to the optimal transforms as K tends to infinity. For a fixed number of vectors K, we will try to evaluate, for both transforms, which bitrate reduction (w.r.t. scheme (1)) is achieved by the corresponding transform. These results are presented in [142]. After the analysis of these single-stage coding schemes, we will move on to two-stages structures based on the KLT and LDU transforms.

- In chapter 7, the integer-to-integer implementation of the two transforms will be further investigated in the framework of figure 5.8. For a fixed preliminary quantization stage (and for sufficiently high resolution), we will analyze the bitrate required to entropy code the low resolution and the error signals. The resulting overall bitrate will be compared to that obtained with the single-stage structures of the previous chapter. We will show that while orthogonal transform tend to “gaussianize” the error signals, the LDU benefits from keeping them uniform. As a consequence, the orthogonal transforms, including the KLT, will be shown to be approximately 0.25 b/s/ch suboptimal w.r.t. their causal counterpart. Finally, we will underline several other practical coding advantages of the LDU, such as the ability of switching easily from a single- to a multi-stage structure, or that of allowing one to quantize with different resolution levels the different channels. These results are presented in [143].

As at the end of chapter 4, we will then generalize the results regarding the causal approach by considering infinite vectors of vector samples, in the frameworks of the single- and multi-stage lossless structures described so far.

- The last chapter deals with optimal lossless coding of vectorial signals. The coding structure investigated in a first step is similar to that of scheme $(\mathit{2})$, or fig. 5.8, where the transform T will be a particular prediction matrix $L(z)$ of the generalized MIMO prediction framework. Similarly to chapter 6, the corresponding compression performance will be compared to the optimal compression performance, as achievable by any lossless coding technique. The particular cases of the classical and the triangular MIMO predictors will be investigated, and shown to present equivalent performance. In a second step, we will generalize the coding scheme of fig. 5.9 by introducing ADPCM loops, whose quantizers allow one to choose the respective bitrates for both the error and the low resolution signals. For these two-stages structures, we will compare, similarly to chap. 7, the overall bitrate delivered by the multiresolution structure to that of the corresponding "one-shot" approach. These two-stages structures will be shown to be slightly suboptimal because of the noise feedback created in ADPCM loops. Finally, the two-stage structure will be generalized to M stages; a strategy will be proposed so that the delivered bitrates approach some predetermined target rates. These results are presented in [144].

Chapter 6

Causal versus Unitary Single-Stage Lossless Transform Coding

In single-stage lossless transform coding, integer-to-integer transforms are used to decorrelate N discrete scalar sources into N transform components. These integer-to-integer implementations involve a cascade of triangular matrices and rounding operations. In [41], the optimality of the integer-to-integer implementation of the Karhunen-Loève Transform (KLT) was established in the limit of negligible round off errors. This chapter presents a similar single-stage, or "one-shot" lossless coding procedure based on the causal LDU transform. We define in a first step the lossless coding *gain for a transformation as the number of bitrate reduction operated by the corresponding lossless coding scheme over a system using no transform. This gain is linked to the mutual information between the random variables (r.v.s) to be coded. In a second step, the effects of the integer-to-integer constraint (round off errors) on the coding gain are analyzed for both the unitary and causal approaches. A third step focuses on the effects of estimation noise on the coding gain: in this case, the transforms are based on a estimate $\widehat{R}_{\underline{x}^q\underline{x}^q}$ of the covariance matrix of the quantized signals $R_{\underline{x}^q\underline{x}^q}$. In any case, the LDU-based approach is shown to yield the highest coding gain. The theoretical analyses are confirmed by numerical results.*

6.1 Introduction

Let us consider the three coding schemes of figure 6.1, simplified from figure 5.8. In all cases, continuous sources x_i are quantized using unbounded uniform scalar quantizers with stepsizes Δ_i (quantization stage Q, def. (7.1)). In the first scheme (1), the resulting discrete valued scalar sources x_i^q [1] are directly entropy coded using a set of independent scalar entropy coders γ_i (codewords i_i with lengths l_{i_i} are transmitted to the decoder).

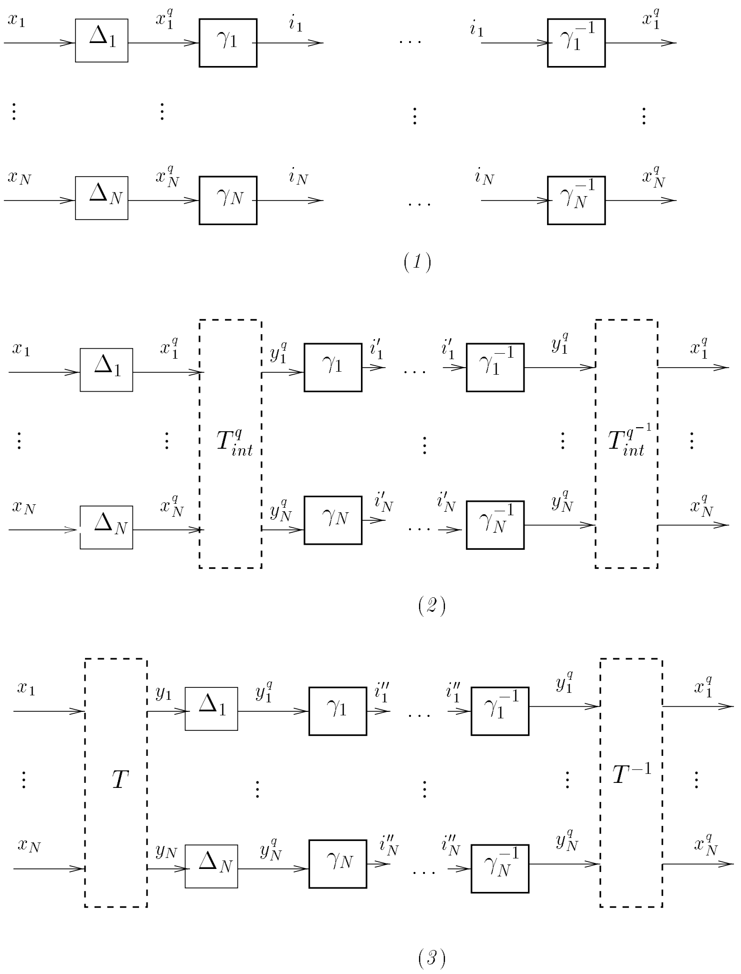


Figure 6.1: Coding schemes considered in this chapter. (1) Direct entropy coding of the x_i^q (2) Introduction of a lossless transform after quantization and (3) Classical transform coding scheme.

As stated in the introduction of this second part, sources of interest x_i may generally present dependencies , and so do indeed their quantized versions. Thus, in order to avoid to code any redundancy, one may apply a transform T_{int}^q, which maps integers to integers, before entropy coding (coding scheme (2)). The

[1] In this chapter, superscript ^{q} will denote quantization in order to emphasize the fact that the sources x_i^q and y_i^q are, up to a scaling factor, integer valued. Subscript $_{int}$ refers to integer-integer implementation of the corresponding transform.

resulting discrete scalar sources y_i^q are further entropy coded (codewords i_i' with lengths $l_{i_i'}$ are transmitted). The transform T_{int}^q is chosen to be invertible so that the decoder can losslessly decode the data x_i^q. Comparing with the classical transform coding framework $(\mathit{3})$, quantization and transformation are reversed. We will refer to the coding scheme $(\mathit{2})$, which from N quantized values produces N discrete transform components, as *single-stage* or *one-shot* lossless coder.

Because the classical transform coding framework $(\mathit{3})$ is very similar to the coding scheme $(\mathit{2})$ in the sense that T_{int}^q is aimed of producing decorrelated transform components, integer-to-integer transforms approximating continuous transforms T have received much attention in the literature [2]. In [41], the framework presented in $(\mathit{2})$ was first introduced as an alternative to transform coding. It is shown that uniform quantization followed by KLT based integer-to-integer mapping and separately encoding of the transform sources is asymptotically (in the limit of high rate, or small Δ_i) as efficient as vector entropy coding the sources x_i^q of scheme $(\mathit{1})$, or scalar entropy coding the components y_i^q of scheme $(\mathit{3})$.

Before these theoretical results, many papers had devised one-to-one integer mapping approaches for simple transforms, such as the S transform [145], the TS transform [146], the S+P [147] and the generalized S transform [148]. The method called "lifting scheme", introduced by Sweldens in [149], was implemented for integer mappings of wavelet transforms in [150] and generalized in [151, 152] and [153]. An integer-to-integer implementation of the DFT is described in [154]. Integer mappings based on lifting steps of (8 point) DCT is exposed in [155], using previous factorizations published by Chen [156] and Loeffler [157] [3]. All these systems are widely used in the framework of lossless image compression. Integer-to-integer tranforms applied to audio coding were compared [21]; the work [159] presents results concerning integer-to integer DWTs to lossless sound compression. Recent work presents general results concerning the factorization (and therefore the integer-to-integer, or "reversible" implementation) of general real-valued transforms, including existence conditions and factorization algorithms [160].

Previous attempts to characterize the performance of integer-to-integer transforms [41, 160] were to find an upper bound for the error induced by a mapping, that is, a bound for $||T_{int}^q(x^q) - Tx^q||_\infty$. In [41], it was shown that for 2×2 unimodular matrices with non-zero coefficients (*e.g.* the KLT), positioned after a quantization stage using equal stepsizes Δ, this bound is

$$||T_{int}^q(x^q) - Tx^q||_\infty \leq (1+K)\frac{\Delta}{2}, \tag{6.1}$$

where $||\underline{x}||_\infty = \max_i |x_i|$, and K is a strictly positive value depending on the coefficients of T. This shows that T_{int}^q precisely approximates T for small stepsizes, and the performance of T_{int}^q and T were proved to be equivalent in the limit of high rate. In this work, we try to go a step further into the analysis of the performance of T_{int}^q by evaluating, in terms of loss in compression, or excess rate, this inherent suboptimality.

First, it may seem natural to define a *lossless coding gain*, which corresponds to the gain, in bits per

[2] Note that we are not interested in building *integer arithmetic* transforms. The computations are still done with floating points numbers, but the result is guaranteed to be integer and invertibility is preserved.

[3] According to [151], the cases of the DFT and the DCT were previously solved by Hong in [158].

sample, obtained by entropy coding the outputs y_i^q of an integer-to-integer transform (scheme $(\mathit{2})$) w.r.t. to that required to entropy code the x_i^q of scheme $(\mathit{1})$. This gain may be thus defined as

$$G_T = \frac{1}{N}\sum_{k=1}^{N} \mathrm{E}_k l_{i_k} - \mathrm{E}_k l_{i'_k} = \frac{1}{N}\sum_{i=1}^{N} H(x_i^q) - H(y_i^q), \tag{6.2}$$

where E_k denotes the expectation over the indexes, and H denotes discrete entropy.

An obvious question is then: given a vectorial source $\underline{x}^q$, obtained from $\underline{x}$ by uniform quantization (Δ_i), what is the maximum bitrate reduction obtained by using scheme $(\mathit{2})$ instead of $(\mathit{1})$? The corresponding gain G_{max}, derived in section 5.1, will then represent an upper bound to the performance of any integer-to-integer transform.

Instead of bounding the errors $||T^q_{int}(x^q) - Tx^q||_\infty$ for both the KLT and the LDU, we will then seek to express in terms of excess rate (or in terms of coding gain reduction w.r.t. G_{max}), the respective integer-to-integer constraints incured by the two transforms. This approach seems natural since minimizing the average bitrate is the most relevant issue in the design of lossless coding systems.

Note that transforms optimized such that the outputs have similar distributions were also presented in [41], allowing one to entropy code these outputs with the same Huffman table, resulting in complexity and memory savings. Further complexity reduction was achieved for Gaussian sources in [161] by using Golomb-Rice instead of Huffman coding. The present work focuses more on the performance of the transformations than on the entropy codes, and their corresponding complexity. The rates of the corresponding transform components will be measured by the discrete entropy, or by those obtained by (multiple-table-based) Huffman coding.

Finally, adaptativity will also be considered, that is, the problem of describing how fast a single-stage compressor which has *a priori* no knowledge about the optimal transform, and whose adaptation is based on the causal past, converges to the optimal performance.

In order to carry a tractable analysis, we will assume a stationary memoryless Gaussian source model, $\underline{x} \sim \mathcal{N}(\mathbf{0}, R_{\underline{x}\underline{x}})$. Moreover, the resolution will be assumed to be sufficiently high, and the p.d.f.s of the signals to be quantized smooth enough, so that quantization with stepsize Δ_i yields uniformly distributed errors (over $[-\frac{\Delta_i}{2}; \frac{\Delta_i}{2}]$), and distortion $\frac{\Delta_i^2}{12}$. We will use the Rényi's relation of differential to discrete entropy for uniformly scalar quantized sources with stepsize Δ_i [38]

$$H(x_i^q) \approx h(x_i) - \log_2 \Delta_i. \tag{6.3}$$

and the similar relation for the N-vectorial source [35, 162]

$$H(\underline{x}^q) + \sum_{i=1}^{N} \log_2 \Delta_i \to h(\underline{x}) \;\text{ as }\; \Delta_i \to 0 \;, i = 1, ..., N. \tag{6.4}$$

In the next section, we derive the expression of the ideal lossless coding gain. The third part compares the causal LDU and unitary KLT approaches to this bound for single-stage lossless coding based on approximation of linear transforms. The fourth section is dedicated to estimation noise and derives the coding gains of the two approaches when the transformations are based on an estimate of the covariance matrix. Finally, the fifth section presents some numerical results.

6.2 Maximum Coding Gain and Mutual Information

6.2.1 Maximum Lossless Coding Gain

The amount of information $H(\underline{x}^q)$ about a vectorial source $\underline{x}^q$ conveyed to the decoder is the same in any lossless coding scheme, either integer-to-integer transform or not. However, a lossless transform coding scheme takes advantage from a non- (or less) redundant repartition of this information among the several signals y_i^q. Assume that these components are made independent by an ideal transform T_{int}. Consider the Venn diagram of figure 6.2. The entropy $H(\underline{x}^q)$ is represented for $N = 2$. In diagram (a), the information conveyed to the decoder is $H(x_1^q) + H(x_2^q) > H(\underline{x}^q)$; in diagram (b), where y_1^q and y_2^q are independent, the vectorial source is represented by $H(y_1^q) + H(y_2^q) = H(\underline{x}^q)$. This intuitively shows that the mutual information between the variables chosen to represent the source should be minimized.

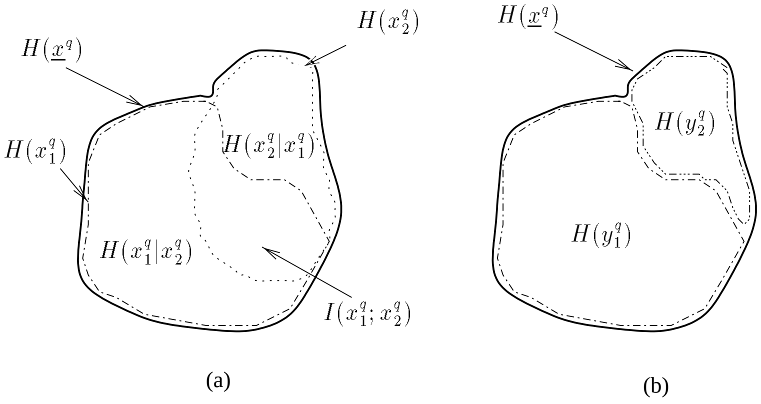


Figure 6.2: Entropy and mutual information

Assume that such a transformation T_{int} exists. If the transform is invertible, the entropy of the vectorial source $\underline{x}^q$ remains unchanged [3], thus the overall bitrate required to independently code the y_i^q is

$$\sum_{i=1}^{N} H(y_i^q) = H(\underline{y}^q) = H(\underline{x}^q), \tag{6.5}$$

which is also the minimum bitrate required to losslessly encode the vectorial source $\underline{x}^q$. These signals y_i^q will then be more suitably scalar entropy coded than the x_i^q. For a Gaussian random variable x_i, the differential entropy $h(x_i)$ equals $\frac{1}{2}\log_2 2\pi e\sigma_{x_i}^2$. It can be easily shown that for sufficiently small quantization stepsizes Δ_i, subsisting (6.4) in (6.5) yield

$$H(\underline{x}^q) = \frac{1}{2}\log_2 \frac{(2\pi e)^N \det R_{\underline{xx}}}{\prod_{i=1}^{N} \Delta_i^2}, \tag{6.6}$$

where $R_{\underline{xx}} = \text{E } \underline{x}\ \underline{x}^T$. The maximum coding gain is then

$$\begin{aligned} G_{max} &= \tfrac{1}{N}\sum_{i=1}^{N} H(x_i^q) - H(\underline{x}^q) \\ &= \tfrac{1}{N}\sum_{i=1}^{N} h(x_i) - h(\underline{x}) \\ &= \tfrac{1}{2N}\log_2 \frac{\det \text{diag}\{R_{\underline{xx}}\}}{\det R_{\underline{xx}}} \\ &= \tfrac{1}{2}\log_2 G_{TC}^{(0)}, \end{aligned} \tag{6.7}$$

where $\text{diag}\,\{R\}$ denotes the diagonal matrix with same diagonal as $\{R\}$, and G_{TC}^0 is the high rate transform coding gain (2.10).

The gain G_{max} is ideal because it corresponds in the Gaussian case to an optimal linear decorrelating transform placed before the quantizers: by writing $\det R_{\underline{xx}} = \prod_{i=1}^{N} \sigma_{y_i}^2 = \prod_{i=1}^{N} \lambda_i$ (where $\sigma_{y_i}^2$ and λ_i are respectively the optimal prediction error variance of x_i based on $\underline{x}_{1:i-1}$, and the eigenvalues of $R_{\underline{xx}}$), we can write equation (6.6) as

$$\begin{aligned} H(\underline{x}^q) &= \sum_{i=1}^{N} \frac{1}{2}\log_2 2\pi e\sigma_{y_i}^2 - \log_2 \Delta_i \\ &= \sum_{i=1}^{N} \frac{1}{2}\log_2 2\pi e\lambda_i - \log_2 \Delta_i, \end{aligned} \tag{6.8}$$

which shows that the entropy of the vector $\underline{x}^q$ may be written as the sum of the entropies of N independent r.v.s of variances $\sigma_{y_i}^2$ (or λ_i), quantized with quantization stepsizes Δ_i. Thus, if we apply a KLT or an LDU to the source $\underline{x}$ before quantization, and then quantize the tranformed signals with stepsizes Δ_i , the minimum bitrate required to entropy code these transformed signals is given by (6.8). Hence, the gain (6.7) would be obtained by a classical transform coding scheme (3).

Another interesting figure in lossless coding is the ratio of the bitrate reduction operated by the lossless coder divided by the bitrate obtained without compression. This *compression ratio* C_{max} is defined as

$$C_{max} = \frac{G_{max}}{\sum_{i=1}^{N} \frac{1}{N} H(x_i^q)} \tag{6.9}$$

As will be illustrated in the next section, the performance of realizable lossless coding schemes based on approximations of linear transforms must be expected to be lower than the expression (6.7): since the transform is placed after the quantizers and just before scalar entropy coders, its output should be discrete valued, which is not the case for optimal linear decorrelating transforms. Thus, rounding operations are necessary; they will induce an entropy increase in the transform signals.

6.2.2 Coding Gains and Mutual Information

A relation of the ideal lossless coding gain (6.7) to the mutual information between quantized r.v.s x_i^q can be obtained as follows.

Let us consider a set of $i-1$ quantized scalar sources x_j^q, $j = 1...i-1$. Assume we wish to code an ith source x_i^q, which is not independent from the $i-1$ others. Intuitively, the best strategy would be to code the only information contained in the ith r.v. which is not shared with the $i-1$ previous variables (cf figure 6.2). The mutual information $I(x_i^q; \underline{x}_{1:i-1}^q)$ allows one to evaluate, loosely speaking, how much information is useless in each r.v., given the knowledge of the other ones. It represents the amount of information that the r.v. x_i^q shares with the $i-1$ others (*i.e.*, the vector $\underline{x}_{1:i-1}^q$), and is defined by

$$I(x_i^q; \underline{x}_{1:i-1}^q) = H(x_i^q) + H(\underline{x}_{1:i-1}^q) - H(x_i^q, \underline{x}_{1:i-1}^q) = H(x_i^q) + H(\underline{x}_{1:i-1}^q) - H(\underline{x}_{1:i}^q). \tag{6.10}$$

By writing the expressions of the mutual information between x_i^q and $\underline{x}_{1:i-1}^q$ for $i = 2, ..., N$, we obtain

$$\begin{aligned} I(x_2^q; x_1^q) &= H(x_2^q) + H(x_1^q) - H(\underline{x}_{1:2}^q) \\ I(x_3^q; \underline{x}_{1:2}^q) &= H(x_3^q) + H(\underline{x}_{1:2}^q) - H(\underline{x}_{1:3}^q) \\ &\vdots \\ I(x_{N-1}^q; \underline{x}_{1:N-2}^q) &= H(x_{N-1}^q) + H(\underline{x}_{1:N-2}^q) - H(\underline{x}_{1:N-1}^q) \\ I(x_N^q; \underline{x}_{1:N-1}^q) &= H(x_N^q) + H(\underline{x}_{1:N-1}^q) - H(\underline{x}_{1:N}^q). \end{aligned} \tag{6.11}$$

Then by summing and averaging the previous expressions, we get

$$\begin{aligned} \frac{1}{N}\sum_{i=2}^{N} I(x_i^q; \underline{x}_{1:i-1}^q) &= \frac{1}{N}\sum_{i=1}^{N} H(x_i^q) - H(\underline{x}^q) \\ &= \frac{1}{N}\sum_{i=1}^{N} h(x) - h(\underline{x}) \\ &= G_{max}. \end{aligned} \tag{6.12}$$

Thus, the maximum bitrate reduction using a lossless transform coding scheme corresponds to the average mutual information shared between each new random variable to be coded and the previous ones. Equivalently, by (6.7), this illustrates why TC is advantageous. By optimally dividing the information $h(\underline{x})$ between the transform components y_i, TC provides w.r.t. to scheme (1) of figure 6.1 a gain G_{TC}^0 for the same rate, or a gain G_{max} in rate for the same distortion. This shows that under high rate assumption and for variable-rate coding, optimal decorrelating transforms such as LDU or KLT may not be optimal for non Gaussian sources since independence, rather than decorrelation, is seeked for [49][4].

Note also from (6.12) that quantizing does not change mutual information, which is correct if the Rényi's relation (6.3) is valid (small Δ_i). This means that the compression ratio C_{max} (eq. 6.9) should increase when the reference rate $\frac{1}{N}\sum_{i=1}^{N} H(x_i^q)$ decreases (see section 6.5).

[4] For fixed-rate coding, there are sources for which even a transform that yields independent components may be suboptimal [46].

6.3 Integer-to-Integer Transforms

We consider N quantized scalar signals x_i^q, which are quantized versions of x_i to the nearest multiple of Δ_i (denoted by $[.]_{\Delta_i}$), and takes values in the set

$$\Delta_i \mathbb{Z} : \underline{x}_{i,k}^q = [x_{1,k}^q, x_{2,k}^q, ..., x_{n,k}^q]^T = [[x_{1,k}]_{\Delta_1}, [x_{2,k}]_{\Delta_2}, ..., [x_{n,k}]_{\Delta_N}]^T .$$

An integer-to-integer transform[5] T_{int}^q: $\Delta_1\mathbb{Z} \times \Delta_2\mathbb{Z} ... \times \Delta_N\mathbb{Z} \rightarrow \Delta_1\mathbb{Z} \times \Delta_2\mathbb{Z} ... \times \Delta_N\mathbb{Z}$ associates to each quantized N-vector $\underline{x}_{i,k}^q$ an N-vector $\underline{y}_{i,k}^q = T\underline{x}_{i,k}^q$ whose components y_i^q are quantized to the same resolution Δ_i as the corresponding x_i^q. The transformation is chosen to be invertible so that the decoder can losslessly compute the original data by $\underline{x}_{i,k}^q = T^{-1}\underline{y}_{i,k}^q$. Since the aim of the transform T_{int}^q is to make the transform signals independent, it can be designed to approximate linear decorrelating transforms such as the LDU or the KLT, which are optimal for Gaussian signals in the classical transform coding case. Although both integer-to-integer implementations tend to the maximum gain of expression (6.7) in the limit of small quantization stepsizes, a quantifiable loss in performance occurs in practical coding situations. This loss is evaluated in the following.

6.3.1 Integer-to-Integer implementation of the LDU

In a first step, the linear transform $L^q = I - \overline{L^q}$ is optimized to decorrelate the quantized data x_i^q. Similarly as in chapter 2, we look for $\min\limits_{L_{i,1:i-1}^q} L_i^q (R_{\underline{x}^q\underline{x}^q}) L_i^{q\,T}$, which leads to the normal equations

$$\begin{bmatrix} \\ R_{\underline{x}^q\underline{x}^q\,1:i,1:i} \\ \\ \end{bmatrix} \begin{bmatrix} L_{i,i-1}^q \\ \vdots \\ L_{i,1}^q \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sigma_{y_i'}^2 \end{bmatrix},$$

where $\sigma_{y_i'}^2$ is the optimal prediction error variance corresponding to the optimal (continuous valued) prediction error $y_{i,k}' = x_{i,k}^q - L_{i,1:i-1}^q \underline{x}_{1:i-1,k}^q = x_{i,k}^q - \widehat{x}_{i,q}^q$. The optimal transform vector is then $\underline{y}'_k = \underline{x}_k^q - \overline{L^q}\,\underline{x}^q$, and the optimal transform L^q corresponds in this case to the LDU factorization of the covariance matrix of quantized data $R_{\underline{x}^q\underline{x}^q} = L^{q\,-1} R_{\underline{y}'\underline{y}'} L^{q\,-T}$. The second step is to design an approximation L_{int}^q of L^q which allows one to keep the transform structure lossless. This can easily be realized by rounding off each estimate $\widehat{x}_{i,q}^q$ of $x_{i,k}^q$. Each transform coefficient is then computed by

$$y_{i,k}^q = x_{i,k}^q - [\widehat{x}_{i,k}^q]_{\Delta_i} = x_{i,k}^q - [L_{i,1:i-1}^q \underline{x}_{1:i-1,k}^q]_{\Delta_i}, \tag{6.13}$$

see figure 6.3.

Let us denote by L^{q_i} the matrix whose non zeros off diagonal elements correspond to the ith optimal

[5] "Integer-to-integer" has be retained in the literature to specify that the transforms are on an integer (but scaled) lattice .

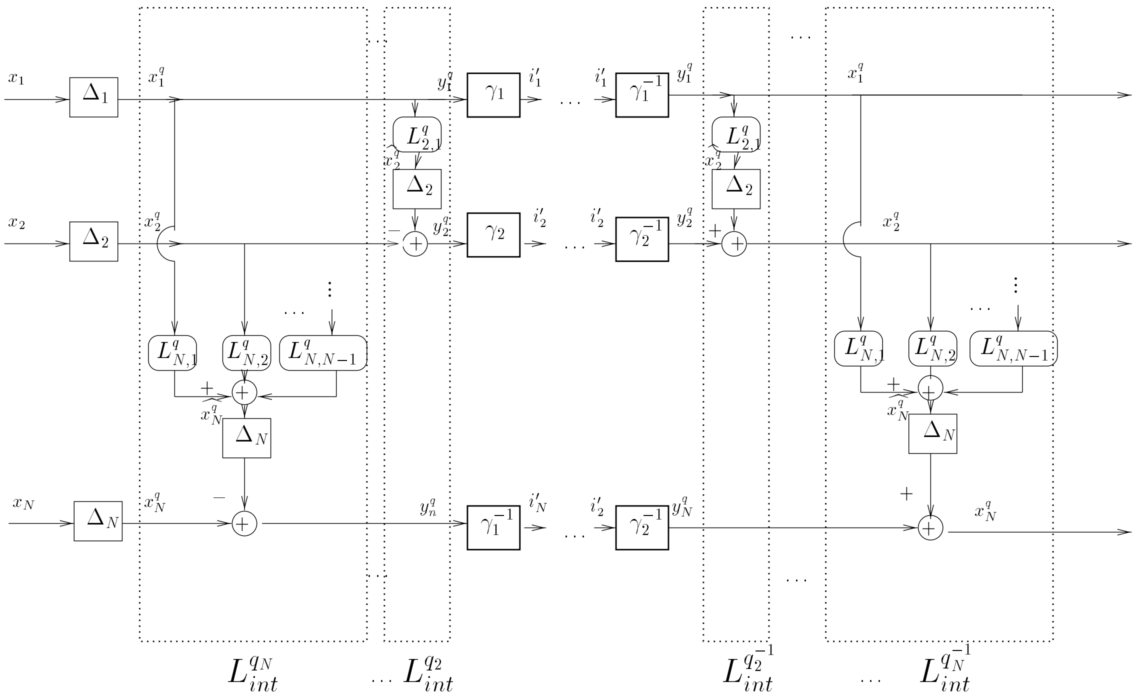


Figure 6.3: Lossless implementation of the LDU transform. An optimal prediction matrix L^q is first computed; the transform coefficients y_i^q are obtained by rounding off and substracting the corresponging estimates $\widehat{x}_i^q$.

predictor [6]

$$L^{q_i} = I - \overline{L}^{q_i} = \begin{bmatrix} 1 & & & & & & \\ 0 & \ddots & & & \mathbf{0} & & \\ \vdots & \ddots & \ddots & & & & \\ 0 & \cdots & 0 & \ddots & & & \\ L_{i,1}^q & \cdots & \cdots & L_{i,i-1}^q & 1 & & \\ 0 & \cdots & \cdots & \cdots & 0 & \ddots & \\ \vdots & & & & & \ddots & \ddots \\ 0 & \cdots & & \cdots & & \cdots & 0 & 1 \end{bmatrix} . \tag{6.14}$$

Then a lossless implementation $L_{int}^{q_i}$ of L^{q_i} is obtained by $\underline{y}_k^q = L_{int}^{q_i}(\underline{x}_k^q) = I - [\overline{L}^{q_i}\,\underline{x}_k^q]_{\Delta_i}$. The inverse operation is simply $\underline{x}_k^q = L_{int}^{q_i^{-1}}(\underline{y}_k^q) = I + [\overline{L}^{q_i}\,\underline{x}_k^q]_{\Delta_i}$.

Now, the global transform vector $\underline{y}_k^q$ can be computed using a cascade of $N-1$ elementary transforms, that

[6] This kind of matrix (called *SERM*, Single-row Elementary Reversible Matrix [160]) appears in many lossless factorizations.

are represented in figure 6.3.

$$\begin{aligned} \underline{y}_k^q &= \left[L^{q_2}\left[L^{q_3}\dots[L^{q_N}\,\underline{x}_k^q]_{\Delta_N}\cdots\right]_{\Delta_3}\right]_{\Delta_2} \\ &= L_{int}^{q_2}\,(L_{int}^{q_3}...(L_{int}^{q_N}\,(\underline{x}_k^q))) \\ &= L_{int}^q(\underline{x}_k^q). \end{aligned} \tag{6.15}$$

At the decoder, the inversion is realized by

$$\begin{aligned} \underline{x}_k^q &= L_{int}^{q^{-1}}\,(\underline{y}_k^q) \\ &= L_{int}^{q_N^{-1}}\,(L_{int}^{q_{N-1}^{-1}}...(L_{int}^{q_2^{-1}}\,(\underline{y}_k^q))). \end{aligned} \tag{6.16}$$

Since the source x_i^q is discrete, we can write the transform components as

$$\begin{aligned} y_{i,k}^q &= x_{i,k}^q - \left[\widehat{x}_{i,k}^q\right]_{\Delta_i} \\ &= \left[x_{i,k}^q - \widehat{x}_{i,k}^q\right]_{\Delta_i} \\ &= \left[y'_{i,k}\right]_{\Delta_i}. \end{aligned} \tag{6.17}$$

This leads to the equivalent representation of L_{int}^q of figure 6.4, where L_{int}^q corresponds to the cascade $\overline{L}^q$ with a quantization stage Q' composed of $N-1$ quantizers. Comparing with eq. (6.1), this shows in particular that for $N=2$, $||L_{int}^q(x^q) - Lx^q||_\infty \leq \frac{\Delta}{2} < ||V_{int}^q(x^q) - Vx^q||_\infty$, meaning that the maximum error is less in the causal than in the unitary case, but this does not give much insight about how the rates are increased.

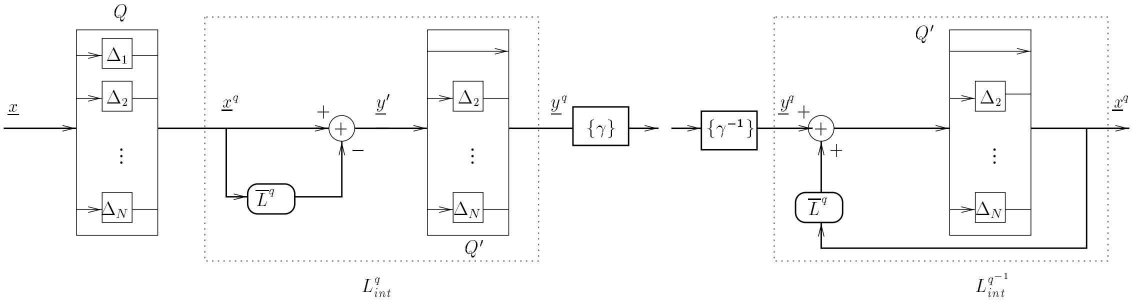


Figure 6.4: Equivalent implementation of the integer-to-integer LDU transform.

We should here underline the similarity between the integer-to-integer implementation of the LDU and the lossless matrixing described in [163], or the decorrelation approach applied to lossless image coding [121]. In these works however, the diagonalizing aspect of the transform (and thus its optimality for Gaussian signals in the case of negligible perturbation effects) was not established. Moreover, the perturbation effects due to the rounding operations (next section) and estimation noise (section 6.4) are not, to our knowledge, analyzed in their published related work.

In order to analyze the effects of the rounding operations, (quantization $[.]_{\Delta_i}$ of the $\widehat{x}_{i,k}^q$) uppon the coding gain, we approximate the entropy $H(y_i^q)$ of the variables y_i^q by $h(y_i')-\log_2\Delta_i$, which assumes a quantization noise uniformly distributed over $[-\frac{\Delta_i}{2},\frac{\Delta_i}{2}]$. The continuous r.v.s y_i' are not strictly Gaussian since each y_i' is a linear combination of i Gaussian r.v.s and $i-1$ uniform r.v.s . Since the p.d.f. of a sum of uniformly distributed r.v.s tends quickly to a Gaussian p.d.f., we assume that this is the case, and this entropy may be evaluated as

$$H(y_i^q)\approx\frac{1}{2}\log_2 2\pi e\sigma_{y_i'}^2-\log_2\Delta_i, \tag{6.18}$$

where $\sigma_{y_i'}^2$ is the actual variance of the ith transform signal. Note that in the integer-to-integer implementation of the LDU, the first scalar signal remains unchanged, and only $N-1$ rounding operations are involved in the lossless transformation. The bitrate required to entropy code the discrete r.v.s y_i^q is then

$$\frac{1}{N}\sum_{i=1}^{N}H(y_i^q)\approx\frac{1}{N}\left(\frac{1}{2}\log_2(2\pi e)\sigma_{x_1}^2-\log_2\Delta_1+\sum_{i=2}^{N}\frac{1}{2}\log_2(2\pi e)\sigma_{y_i'}^2-\log_2\Delta_i\right). \tag{6.19}$$

The lossless coding gain for the integer-to-integer LDU may then be written as

$$\begin{aligned} G_{L_{int}^q} &= \frac{1}{N}\sum_{i=1}^{N}H(x_i^q)-H(y_i^q)\\ &\approx \frac{1}{2N}\log_2\frac{\prod_{i=2}^{N}\sigma_{x_i}^2}{\prod_{i=2}^{N}\sigma_{y_i'}^2}\approx\frac{1}{2N}\log_2\frac{\det\text{diag}\{R_{\underline{x}\underline{x}}\}}{\sigma^2{}_{x_1}\prod_{i=2}^{N}\sigma_{y_i'}^2}, \end{aligned}\tag{6.20}$$

The last equality shows that $G_{L_{int}^q}$ is indeed inferior to G_{max} since the denominator involves the optimal prediction error variances obtained from $R_{\underline{x}^q\underline{x}^q}=R_{\underline{x}\underline{x}}+D$ (where D is the diagonal matrix of the distortions, $D_{ii}\approx\Delta_i^2/12$), instead of $R_{\underline{x}\underline{x}}$.

Moreover, since L^q diagonalizes $R_{\underline{x}^q\underline{x}^q}$, we have $\prod_{i=1}^{N}\sigma_{y_i'}^2=\det R_{\underline{x}^q\underline{x}^q}$, where $\sigma_{y_1'}^2=\sigma_{x_1^q}^2$. Using the last equality, the coding gain $G_{L_{int}^q}$ may alternatively be approximated as

$$\begin{aligned} G_{L_{int}^q} &\approx \frac{1}{2}\log_2\frac{\det\text{diag}\{R_{\underline{x}\underline{x}}\}}{\left(\sigma_{x_1^q}^2-\frac{\Delta_1^2}{12}\right)\prod_{i=1}^{N}\sigma_{y_i'}^2}\\ &\approx \frac{1}{2}\log_2\frac{\det\text{diag}\{R_{\underline{x}\underline{x}}\}}{\det R_{\underline{x}^q\underline{x}^q}}+\frac{1}{2}\log_2\left(1+\frac{\Delta_1^2}{12\sigma_{x_1^q}^2}\right)\\ &\approx \frac{1}{2}\log_2\frac{\det\text{diag}\{R_{\underline{x}\underline{x}}\}}{\det R_{\underline{x}^q\underline{x}^q}}+\frac{\Delta_1^2}{24\ln 2\sigma_{x_1}^2}\\ G_{L_{int}^q} &\approx G_{max}-\frac{1}{2N\ln 2}\text{tr}\{DR_{\underline{x}\underline{x}}^{-1}\}+\frac{\Delta_1^2}{24\ln 2\sigma_{x_1}^2}, \end{aligned}\tag{6.21}$$

which clearly expresses the loss due to the lossless constraint w.r.t. the optimal performance. This last expression shows that one should position the most coarsely quantized signal (highest $\frac{\Delta_i}{\sigma_{x_i}}$) in first position in order to maximize $G_{L_{int}^q}$ (see section 6.5.1). Moreover, one can check that $G_{L_{int}^q}$ tends to G_{max} as

$\Delta_i \to 0$, $i = 1, ..., N$, which means that the transform is optimal in terms of lossless coding gains in the case of negligible rounding effects.

6.3.2 Integer-to-Integer implementation of the KLT

As far as the KLT is concerned, the integer-to-integer approximation is based on the factorization of a unimodular matrix cascaded with roundings ensuring the inversibility of the global transform. In [41], this transform was shown to be equivalent to the original KLT for arbitrarily small Δ_i. The loss in compression due to the rounding operations is evaluated here in the $N = 2$ case.

Let us denote by V^q a KLT of $R_{\underline{x}^q \underline{x}^q}$. Then we have

$$\Lambda^q = V^q R_{\underline{x}^q \underline{x}^q} V^{qT}, \tag{6.22}$$

and we denote by λ_i^q the variances of the (real-valued) transform signals.

We recall now the construction of the integer-to-integer transform based on V^q. As any unimodular transform with nonzero coefficients, V^q can be factored into three unit diagonal triangular matrices with unit diagonal as

$$\begin{aligned} V^q &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} = V_1^q \, V_2^q \, V_3^q, \\ V_1^q &= \begin{bmatrix} 1 & \frac{a-1}{c} \\ 0 & 1 \end{bmatrix}, \; V_2^q = \begin{bmatrix} 1 & 0 \\ c & 1 \end{bmatrix}, V_3^q = \begin{bmatrix} 1 & \frac{d-1}{c} \\ 0 & 1 \end{bmatrix}. \end{aligned} \tag{6.23}$$

The transform vector $\underline{y}_k$ is then losslessly obtained by using the three-step integer-to-integer transform V_{int}^q

$$\underline{y}_k^q = V_{int}^q \underline{x}_k^q = \left[\underbrace{V_1^q \left[\underbrace{V_2^q \left[\underbrace{V_3^q \, \underline{x}_k^q}_{\underline{y}_k^1} \right]_{\Delta_1}}_{\underline{y}_k^2} \right]_{\Delta_2}}_{\underline{y}_k^3} \right]_{\Delta_1}. \tag{6.24}$$

Since the matrices are triangular, their inverses are simply computed by changing the signs of the off-diagonal elements.

One can analyze the effects of the roundings at each step. Denoting by $\delta_{i,j}$ the error caused by rounding the ith component of the vector $\underline{y}_k^j$, it can easily be shown that the final (discrete valued) transform vector $\underline{y}_k^q$ is obtained by

$$\underline{y}_k^q = \begin{bmatrix} y_{1,k} \\ y_{2,k} \end{bmatrix} = \begin{bmatrix} \left[x_1^q + \frac{d-1}{c} x_2^q + \delta_{1,1} + \frac{a-1}{c}(c x_1^q + c\delta_{1,1} + d x_2^q + \delta_{2,2}) \right]_{\Delta_1} \\ [c x_1^q + c\delta_{1,1} + d x_2^q]_{\Delta_2} \end{bmatrix}. \tag{6.25}$$

Assuming small quantization stepsizes (ensuring the independence of the quantization noises $\delta_{i,j}$), and Gaussianity for the transformed signals, the discrete entropy of each transformed random variable may be approximated as

$$\begin{aligned} H(y_1^q) &\approx \tfrac{1}{2}\log_2 2\pi e(\underbrace{\lambda_1^q + \frac{a^2\Delta_1^2}{12} + \frac{(a-1)^2}{c}\frac{\Delta_1^2}{12}}_{\lambda_1^{'q}=\lambda_1^q+e_1}) - \log_2\Delta_1 \\ H(y_2^q) &\approx \tfrac{1}{2}\log_2 2\pi e(\underbrace{\lambda_2^q + \frac{c^2\Delta_1^2}{12}}_{\lambda_2^{'q}=\lambda_2^q+e_2}) - \log_2\Delta_2. \end{aligned} \tag{6.26}$$

Thus, y_i^q may be seen as a continuous r.v. of variance $\lambda_i^{'q} = \lambda_i^q + e_i$, quantized with stepsize Δ_i. The terms e_i are the increase in the variance of the transform signals due to the rounding operations. The corresponding expression for the lossless coding gain in the $N = 2$ case is then

$$\begin{aligned} G_{V_{int}^q} &= \tfrac{1}{N}\sum_{i=1}^{2} H(x_i^q) - H(y_i^q) \\ &= \tfrac{1}{2N}\log_2 \frac{\prod_{i=1}^{2}\sigma_{x_i}^2}{\prod_{i=1}^{2}\lambda_i^{'q}}, \end{aligned} \tag{6.27}$$

Comparing with the gain obtained for the lossless implementation of the LDU (6.20) we have $G_{V^q,int} < G_{L^q,int}$ (this comes from the following series of inequalities $\prod_{i=1}^{2}\lambda_i^{'q} > \prod_{i=1}^{2}\lambda_i^q = \prod_{i=1}^{2}\sigma_{y_i'}^2 > \sigma_{x_i}^2\sigma_{y_2'}^2$). Thus the gain for the integer-to-integer KLT is clearly inferior to that of the integer-to-integer LDU for the $N = 2$ case. Indeed, only one rounding is used in the LDU case for $N = 2$, whereas three roundings are necessary to losslessly implement the KLT. In the general N case, the triangular structure of the prediction matrix allows one to implement the lossless causal transform using $N - 1$ rounding operations (see (6.15)), which is most probably less than the number required in the unitary case, where the transform matrix has not a triangular structure [7].

An alternative expression of $G_{V_{int}^q}$ may be obtained by approximating the following product under high resolution assumption

$$\prod_{i=1}^{2}\lambda_i^{'q} = \prod_{i=1}^{2}\lambda_i^q\left(1+\frac{e_i}{\lambda_i^q}\right) \approx \prod_{i=1}^{2}\lambda_i^q\left(1+\sum_{i=1}^{2}\frac{e_i}{\lambda_i^q}\right). \tag{6.28}$$

[7] By [160], Th.4, Corol.6, an $N \times N$ orthogonal transform may be factorized as $N + 1$ SERM (and a permutation matrix) of the form (6.14)

We get

$$\begin{aligned}
G_{V^q_{int}} &\approx \tfrac{1}{2}\log_2 \frac{\det\ \text{diag}\,\{R_{\underline{x}\underline{x}}\}}{\det R_{\underline{x}^q\underline{x}^q}} - \tfrac{1}{2N}\log_2\left(1+\sum_{i=1}^{2}\frac{e_i}{\lambda_i^q}\right) \\
&\approx \tfrac{1}{2}\log_2 \frac{\det\ \text{diag}\,\{R_{\underline{x}\underline{x}}\}}{\det R_{\underline{x}^q\underline{x}^q}} - \tfrac{1}{2N\ln 2}\sum_{i=1}^{2}\frac{e_i}{\lambda_i^q} \\
G_{V^q_{int}} &\approx G_{max} - \tfrac{1}{2N\ln 2}\left(\text{tr}\,\{DR^{-1}_{\underline{x}\underline{x}}\} + \sum_{i=1}^{2}\frac{e_i}{\lambda_i^q}\right).
\end{aligned} \tag{6.29}$$

As (6.27), this expression holds for $N = 2$, since the perturbation terms on the variances e_i in (6.26) have been analytically derived in this case only. However, the effects of the rounding can be similarly evaluated for a general N, and the expressions (6.27), (6.29) would hold more generally by plugging in the corresponding e_i. Finally, as expected, $G_{V^q_{int}}$ tends to G_{max} as Δ_i tends to 0, $i = 1, ..., N$.

6.4 Adaptive Systems: Effects of the Estimation Noise

In the vein of chapters 3 and 4, the following analysis focuses on the lossless coding gains of an adaptive scheme based on an estimate of the covariance matrix

$$\widehat{R}_{\underline{x}^q\underline{x}^q} = R_{\underline{x}^q\underline{x}^q} + \Delta R = \frac{1}{K}\sum_{k=1}^{K}\underline{x}^q_k\underline{x}^{qT}_k, \tag{6.30}$$

where K is the number of previously decoded vector available at the decoder. We suppose independent identically distributed Gaussian real vectors $\underline{x}^q_k$ (again, the r.v.s are not strictly Gaussian because of the contribution of the uniform quantization noise; this contribution is however small for a high resolution quantization). In this case, the first and second order statistics of ΔR may be analytically evaluated (see sections 3.A and 3.B): $(\Delta R)_{ii}$ may, for sufficiently large K, be approximated as a zero mean Gaussian random variable with covariance matrix such that $\text{E}\,\text{vec}(\Delta R)\,(\text{vec}(\Delta R))^T \approx \frac{2}{K}R_{\underline{x}^q\underline{x}^q} \otimes R_{\underline{x}^q\underline{x}^q}$, where $\otimes$ denotes the Kronecker product. For each realization of ΔR, the coder computes a in a first step the linear transformation $\widehat{T}$ ($\widehat{T} = \widehat{L}^q$ or $\widehat{V}^q$) which diagonalizes $\widehat{R}_{\underline{x}^q\underline{x}^q}$: $\widehat{T}\widehat{R}_{\underline{x}^q\underline{x}^q}\widehat{T} = \widehat{\Sigma}$. Then, by using the lossless factorizations of the previous sections, the encoder computes the corresponding integer-to-integer transform $\widehat{T}_{int}$. The coding gain $G_{\widehat{T}_{int}}(K)$ is then the expected bitrate reduction w.r.t. to a scheme without transform, for a transform based on K vectors. Equivalently, this is the expected gain obtained for a scheme which stops adapting the transform after K vectors, asympotically in the data length. We assume that the entropy coder possesses N universal lossless codes for the N transform coefficients streams.

6.4.1 Coding Gain for the integer-to-integer LDU

The coding gain is in the causal case

$$G_{\widehat{L}^q_{int}}(K) = \sum_{i=1}^{N} H(x^q_i) - H(y^q_i, K), \tag{6.31}$$

where only the entropies $H(y_i^q, K)$ of the discrete variables y_i^q, obtained by applying $\widehat{L}_{int}^q$ to $\underline{x}^q$, depend on K. Since the variance of the first variable y_1^q is not affected by the transformation, we have

$$H(y_1^q, K) = H(x_1^q) = \frac{1}{2}\log_2(2\pi e)\sigma_{x_1}^2 - \log_2 \Delta_1. \tag{6.32}$$

Concerning the $N-1$ remaining r.v.s y_i^q, they may be seen as r.v.s obtained by applying L^q to $\underline{x}^q$, and then by quantizing the continuous valued result with stepsize Δ_i. Thus, by denoting $(\widehat{L^q} R_{\underline{x}^q\underline{x}^q} \widehat{L^q})_{ii} = (R_{\underline{y}'\underline{y}'})_{ii} + \Delta(R_{\underline{y}'\underline{y}'})_{ii}$, we obtain

$$\begin{aligned} H(y_i^q, K) &= \mathrm{E}\ \tfrac{1}{2}\log_2(2\pi e)(\widehat{L^q} R_{\underline{x}^q\underline{x}^q} \widehat{L^q})_{ii} - \log_2 \Delta_i \\ &= \mathrm{E}\,\tfrac{1}{2}\log_2\left(2\pi e(R_{\underline{y}'\underline{y}'})_{ii}\left(1+\tfrac{\Delta(R_{\underline{y}'\underline{y}'})_{ii}}{(R_{\underline{y}'\underline{y}'})_{ii}}\right)\right) - \log_2 \Delta_i \\ &\approx \tfrac{1}{2}\log_2 2\pi e(R_{\underline{y}'\underline{y}'})_{ii} - \log_2 \Delta_i + \tfrac{1}{2\ln 2}\,\mathrm{E}\ \tfrac{\Delta(R_{\underline{y}'\underline{y}'})_{ii}}{(R_{\underline{y}'\underline{y}'})_{ii}}. \end{aligned} \tag{6.33}$$

Therefore,

$$\begin{aligned} \sum_{i=1}^{N} H(y_i^q, K) &= \tfrac{1}{2}\log_2(2\pi e)\sigma_{x_1}^2 - \log_2 \Delta_1 + \sum_{i=2}^{N}\frac{1}{2}\log_2(2\pi e)^{N-1}\sigma_{y_i'}^2 \\ &\quad - \log_2 \Delta_i + \sum_{i=2}^{N}\frac{1}{2\ln 2}\,\mathrm{E}\ \frac{\Delta(R_{\underline{y}'\underline{y}'})_{ii}}{(R_{\underline{y}'\underline{y}'})_{ii}}. \end{aligned} \tag{6.34}$$

Comparing with the bitrate required to code the y_i^q when the transformation is not perturbed (6.19), the last term corresponds to an excess bitrate due to estimation noise. Using the fact that $\mathrm{E}\ \Delta(R_{\underline{y}'\underline{y}'})_{11} = 0$, this term may be written as

$$\sum_{i=2}^{N}\frac{1}{2\ln 2}\,\mathrm{E}\ \frac{\Delta(R_{\underline{y}'\underline{y}'})_{ii}}{(R_{\underline{y}'\underline{y}'})_{ii}} = \frac{1}{2\ln 2}\,\mathrm{E}\ \sum_{i=1}^{N}\frac{\Delta(R_{\underline{y}'\underline{y}'})_{ii}}{(R_{\underline{y}'\underline{y}'})_{ii}} \approx \frac{N(N-1)}{4\ln 2K}. \tag{6.35}$$

Finally, the lossless coding gain for an integer-to-integer implementation of the LDU when the transform is based on K observed vectors may be approximated as

$$\begin{aligned} G_{\widehat{L}_{int}^q}(K) &= \tfrac{1}{N}\sum_{i=1}^{N} H(x_i^q) - H(y_i^q, K) \\ &\approx G_{L_{int}^q} - \tfrac{N-1}{4\ln 2K}. \end{aligned} \tag{6.36}$$

for large K and under high resolution assumption.

6.4.2 Lossless Coding Gain for the integer-to-integer KLT

In this case, one has to compute the difference

$$G_{\widehat{V}_{int}^q}(K) = \frac{1}{N}\sum_{i=1}^{N} H(x_i^q) - H(y_i^q, K), \tag{6.37}$$

where only the entropies $H(y_i^q, K)$ of the discrete variables y_i^q, obtained by applying $\widehat{V}_{int}^q$ to $\underline{x}^q$, depend on K.

Using a similar analysis, the lossless coding gain with estimation noise for the integer-to-integer KLT may be approximated as

$$\begin{aligned} G_{\widehat{V}^q_{int}}(K) &\approx \tfrac{1}{2}\log_2 \frac{\det \text{diag}\{R_{\underline{x}\underline{x}}\}}{\det R_{\underline{x}^q\underline{x}^q}} - \frac{1}{2\ln 2}\sum_{i=1}^{N}\frac{e_i}{\lambda_i^q} - \frac{N(N-1)}{4\ln 2K} \\ &\approx G_{V^q_{int}} - \tfrac{N-1}{4\ln 2K}, \end{aligned} \tag{6.38}$$

under high resolution assumption and for sufficiently high K. As in section 6.3.2, this expression holds for $N = 2$ (in which case we have derived analytically the gain $G_{V^q_{int}}$), but would hold more generally with the corresponding $G_{V^q_{int}}$.

6.5 Numerical Examples

In the first part of this section, we compare the lossless coding gains obtained for the integer-to-integer implementations of the LDU and the KLT for N=2. Then simulations results for higher values of N are presented in the case of the LDU. The second part of this section describes the effects of estimation noise on the coding gains. We used either entropy or Huffman coded uniform scalar quantizers, and real Gaussian i.i.d. vectors.

6.5.1 Lossless Coding Gains without Estimation Noise

In order to check the theoretical results we generated real Gaussian vectors of covariance matrix $R_{\underline{x}\underline{x}}$ (covariance matrix of a first order autoregressive process with normalized correlation coefficient $\rho = 0.9$). The number of vectors was $N_0 = 10^4$. The vectors were quantized using the same normalized quantization stepsize $\frac{\Delta_i}{\sigma_{x_i}}$. For several values of $\frac{\Delta}{\sigma_x}$, the optimal decorrelating transformations L^q and V^q were computed using the covariance matrix $R^{(N_0)}_{\underline{x}^q\underline{x}^q}$ of the whole data set, that is, $R^{(N_0)}_{\underline{x}^q\underline{x}^q} = \frac{1}{N_0}\sum_{i=1}^{N_0} \underline{x}_i^q \underline{x}_i^{q^T}$. The integer-to-integer transforms L^q_{int} and V^q_{int}, based on the transforms L^q and V^q were implemented and used to compute the transformed data $\underline{y}_i^q$. We repeated this experiment ten times and averaged the different obtained gains.

Results for $N = 2$.

The theoretical maximum coding gain is related to the mutual information between the unquantized variables as expressed in (6.12). The theoretical gains for LDU and KLT are then given by (6.21) and (6.29) respectively. The observed lossless coding gains were then computed in three different ways. Firstly, by computing the 0th order entropies of the discrete transform signals. Secondly by measuring the average length obtained with Huffman codes. Under high resolution assumption the Rényi relation asumes a one to one correspondence between the discrete entropy and the variances of the transform signals through the relations (6.18) for the LDU and (6.26) for the KLT. Thus finally, a third way is to measure the actual variances of the transform signals before quantization ($\sigma^2_{y'_i}$ and $\lambda_i^{q'}$ for the KLT). This allows one to check if the analysis concerning the variances is accurate. In this case, the observed gains are obtained by computing (6.21) and (6.29) with the measured variances of the transform signals. These gains are denoted by "Observed Gain Transform Var." These gains are plotted in figure 6.5 versus $\frac{\Delta}{\sigma_x}$. For high resolution (small values of $\frac{\Delta}{\sigma_x}$), there is a good match between the observed gains and the analytical expressions. In particular, it can be seen from the estimates of the gains based on the actual variances and on the entropies, that the assumptions of Gaussianity and of high resolution are fairly precise for values of $\frac{\Delta}{\sigma}$ less than approximately 0.8. The bitrate reduction obtained by using integer-to-integer transforms is not negligible, even for $N = 2$. Figure 6.6 illustrates the compression ratio of the two analyzed integer-to-integer transform. The maximum achievable compression ratio C_{max} is given by (6.9). Basing our observations on the rates obtained with Huffman codes, a compression ratio of 11% can be operated for $\frac{\Delta}{\sigma_x} = 0.1$ by using any of the two integer-to-integer transformations analyzed in this work. For $\frac{\Delta}{\sigma_x} = 0.51$, the compression ratio is 16% for the integer-to-integer implementation of the LDU, and 14% for the integer-to-integer implementation of the KLT. (For higher values of N, higher compression ratios can be achived). Also, note that high compression ratios are still achievable in the case of coarse quantization.

Considering again figure 6.5, the rounding effects due to the lossless implementation of the transforms indeed can be seen to increase as the quantization gets more coarse. The observed coding gains based on the estimates of the variances of the transformed signals correspond well to the predicted ones until a ratio $\frac{\Delta}{\sigma_x} \approx 1$. When the quantization becomes even more coarse, the quantization noises are not independent anymore, and the mutual information between the quantized variables x_i^q is superior to the theoretical one. Figure 6.7 shows the normalized correlation coefficients of the quantization noise versus the normalized correlation coefficient of the variables x_1 and x_2 for several quantization stepsizes. It indicates that for most of quantization situations, the hypothesis of independence of the quantization noises is reasonnable. When the correlation is not negligible, the transforms take more advantage of the information shared between the quantized variables, and the gains may become superior to the predicted ones. The curves obtained for $N = 2$ are well matched by the theoretical analysis for $\frac{\Delta}{\sigma_x}$ lower than approximately 0.8. They show that a noticeable part of the bitrate may be saved by using an integer-to-integer transform. Finally, the lossless implementation of the LDU provides better performance than that of the KLT.

Position of the first signal

Figure 6.8 shows the codings gains obtained for the integer-to-integer LDU applied to scalar sources of unit variance, versus their correlation coefficient ρ. In the first case, denoted by "1" in the legend, the first signal x_1 is quantized with stepsize $\Delta_1 = 0.1$ and the second signal x_2 with stepsize $\Delta_1 = 1$. In the second case, denoted by "2" in the legend, the stepsizes are 1 for x_1 and 0.1 for x_2. The curves show, as expected, that the most coarsely quantized signal must be placed in first place in order to maximize the lossless coding gain.

Results for $N > 2$.

The coding gains (estimates based on measured variances and Huffman codes) obtained for the integer-to-integer LDU with $N = 5, \Delta = 0.51$ and $N = 5, \Delta = 0.21$ are presented in figures 6.9 and 6.10 respectively. In this case, the data were composed of real Gaussian i.i.d. vectors with covariance matrix $R = HR_{\underline{xx}}H^T$. $R_{\underline{xx}}$ is the covariance matrix of a first order autoregressive process with normalized correlation coefficient ρ. H is a diagonal matrix whose ith entry is $(i)^{2/3}$ (increasing variances, ranging from 1 to 8.56). Hence, the coarseness of the quantization decreases as i increases. It can be seen that the observed gains match well the predicted ones. Figure 6.11 (resp. 6.12) compares the compression ratio (resp. the lossless coding gain) versus N for several values of $\frac{\Delta}{\sigma}$ for the LDU. Note that whereas the theoretical coding gain does not depend on the quantization (the mutual information is theoretically the same between unquantized and quantized r.v.s by 6.12), the compression ratio (percentage of the bitrate reduction caused by the transform w.r.t. to the overall bitrate of the uncompressed data) does. For fine quantization, the maximum compression ratio is relatively low, but may be achieved by an integer-to-integer transform because the effects of the roundings are not too strong. When the quantization becomes more coarse, better compression may be achieved, but on the other hand, the integer-to-integer constraint moves the actual performance of the transfom away from the the optimal performance. However, it appears from this figure that is always advantageous to use an integer-to-integer transform, even in cases of coarse quantization.

6.5.2 Coding Gains with Estimation Noise

In the first experiment, $N = 2$. The coding gains with estimation noise are plotted in figure 6.13. The normalized quantization stepsize is $\frac{\Delta}{\sigma_x} = 0.51$. The coding gain G_{max} refers to the mutual information given by (6.12). The theoretical gains for LDU and KLT are given by (6.36) and (6.38) respectively (gains referred to as “G(K) Transform Asymptotic”). The observed coding gains are either based on the estimates of the variances of the transform signals (gains referred to as “G(K) observed variances”), or based on the actual gain computed by Huffman coding. In this case, a Huffman code is designed for the signals obtained with integer-to-integer transforms based on an estimate of the covariance matrix of quantized data $\widehat{R}_{\underline{x}^q\underline{x}^q}$ with K vectors. The theoretical curves correspond well to the observed ones for the observed gains based on variances estimates for $K \approx$ a few tens. Huffman based and variance based observed gains reach 90% of their maximal value for $K \approx 10$ decoded vectors. That is, regarding the results obtained with Huffman codes, 90% of an optimal compression of 16% can be achieved for $K \approx 10$ in the case of the integer-to-integer LDU. In the case of the integer-to-integer KLT, 90% of a compression of 14% can be achieved for a comparable estimation noise.

Finally, figures 6.14 (resp. 6.15) plot the lossless coding gain with estimation noise versus K for $N = 5$ and $\Delta = 0.51$ (resp. $\Delta = 0.21$). Theoretical and observed gains correspond well for $K \approx$ a few tens.

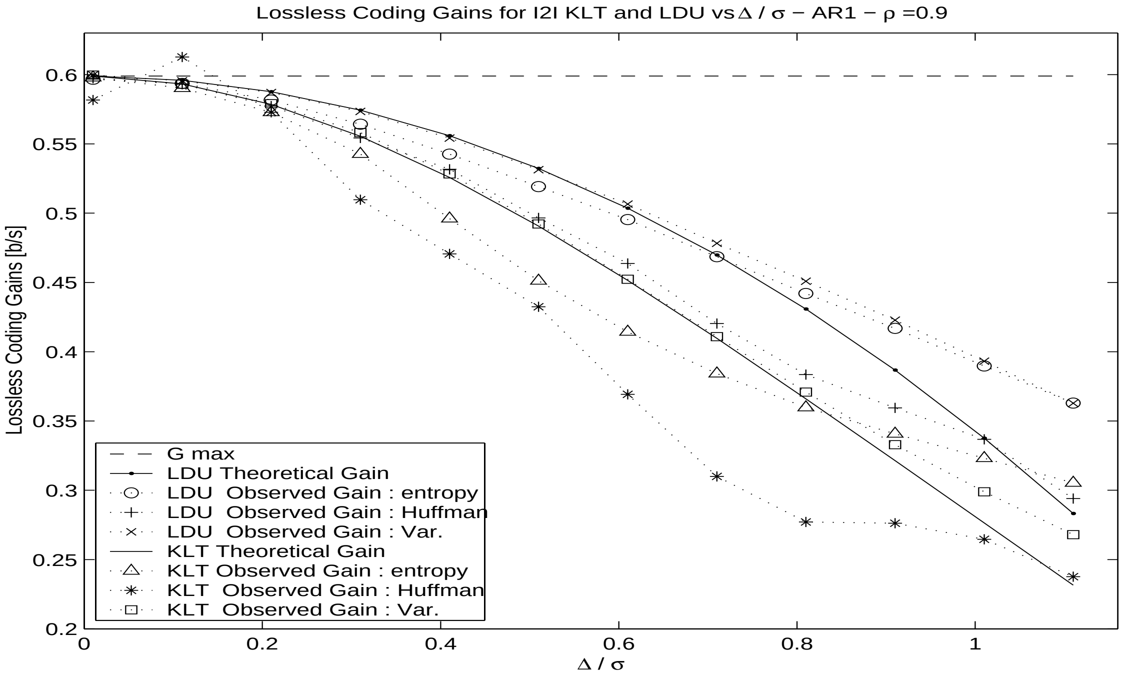


Figure 6.5: Lossless coding gains for integer-to-integer implementations of the LDU and KLT *vs* quantization stepsize. $N = 2$ and $\rho = 0.9$.

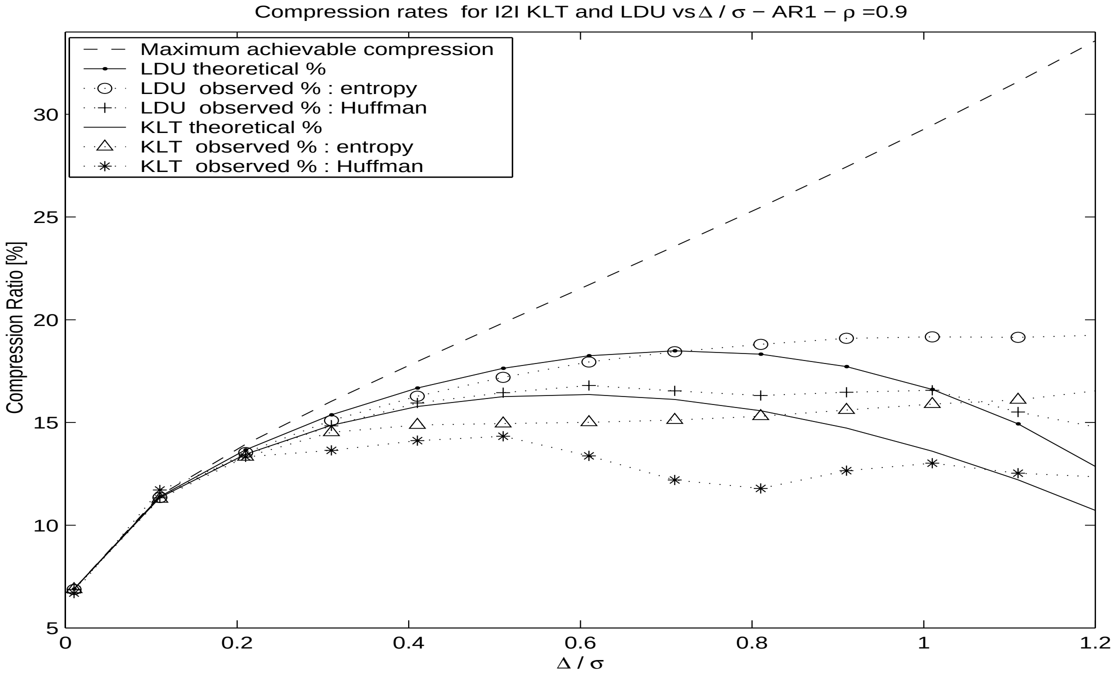


Figure 6.6: Compression ratios achieved by integer-to-integer transforms. $N = 2$ and $\rho = 0.9$.

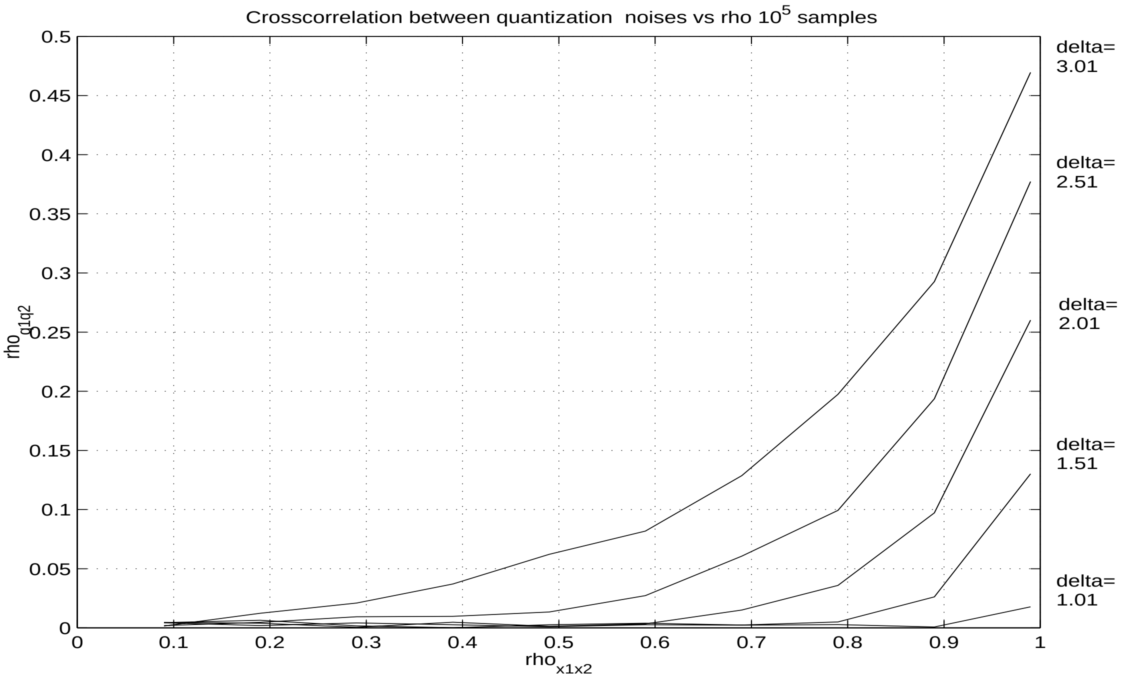


Figure 6.7: Correlation coefficient of quantization noises versus correlation coefficient of the variables x_1 and x_2 for several quantization stepsizes. The variables x_1 and x_2 have variance 1.

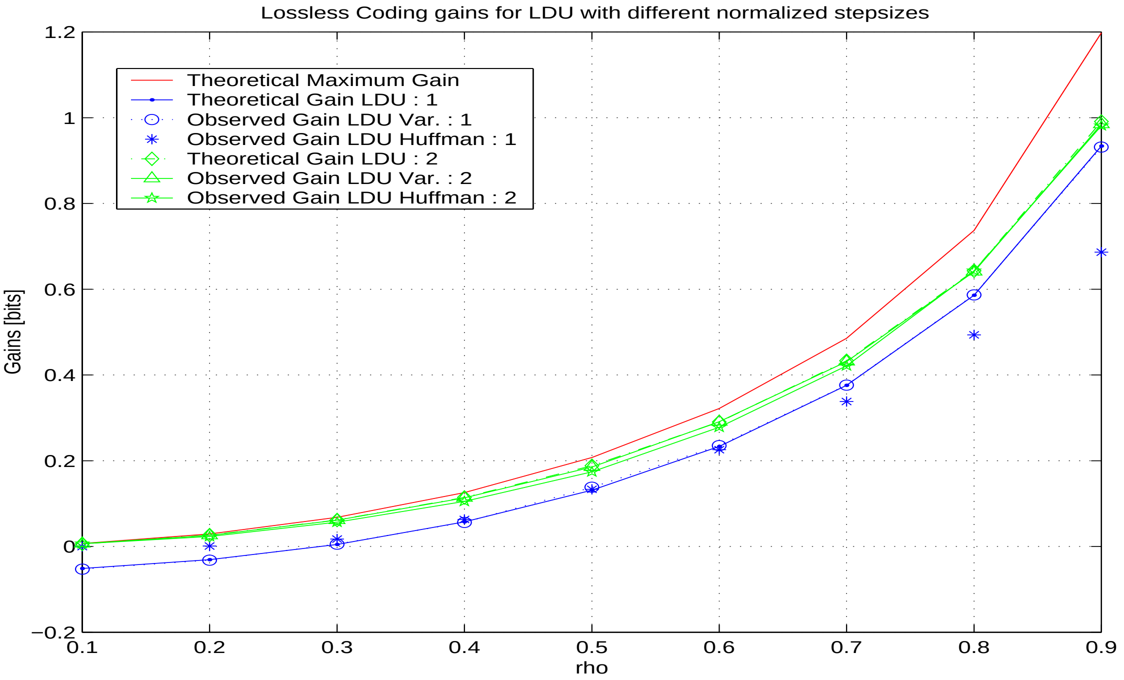


Figure 6.8: Importance of the coarseness of the quantization of the first signal.

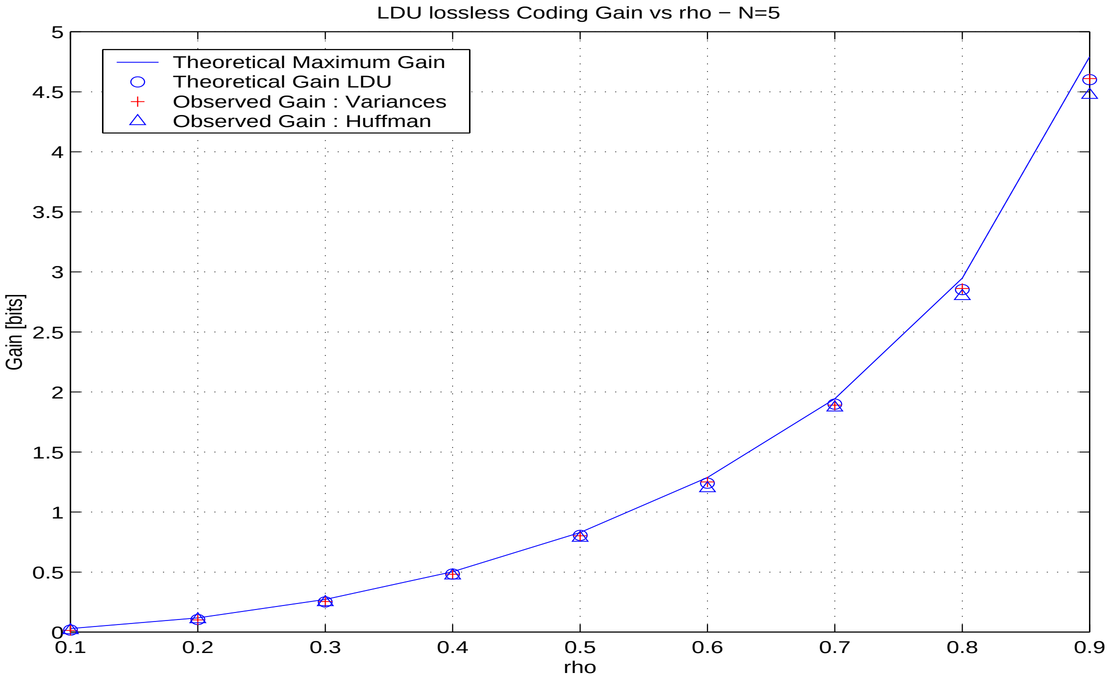


Figure 6.9: Lossless coding gain for integer-to-integer LDU with $N = 5$. $\Delta = 0.51$.

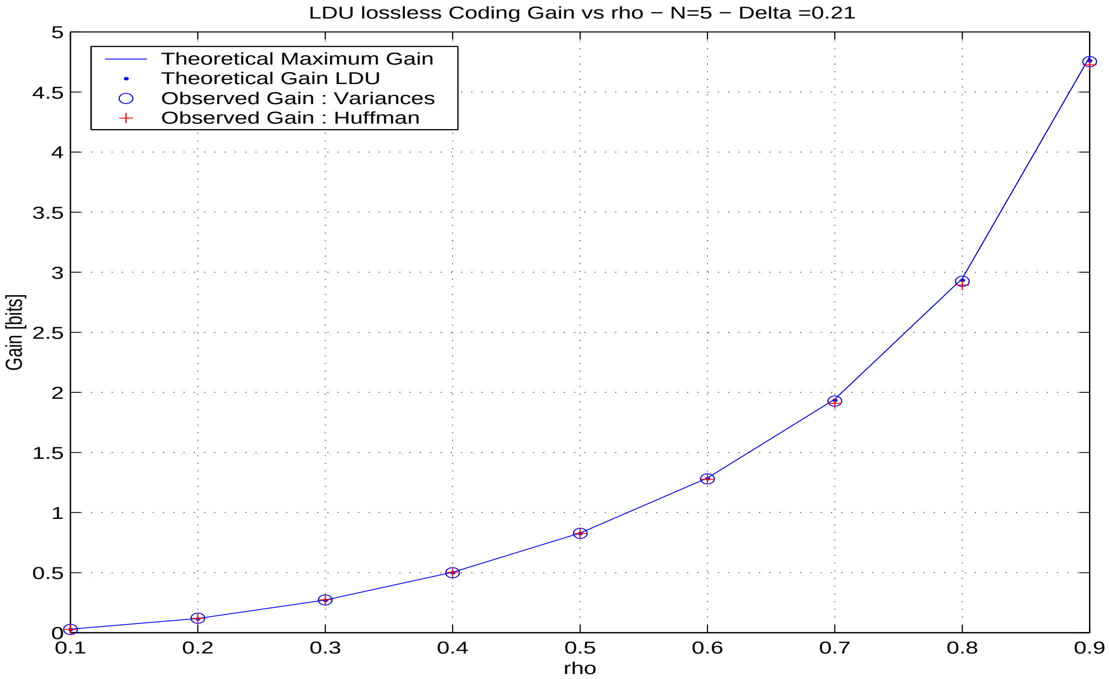


Figure 6.10: Lossless coding gain for integer-to-integer LDU with $N = 5$. $\Delta = 0.21$.

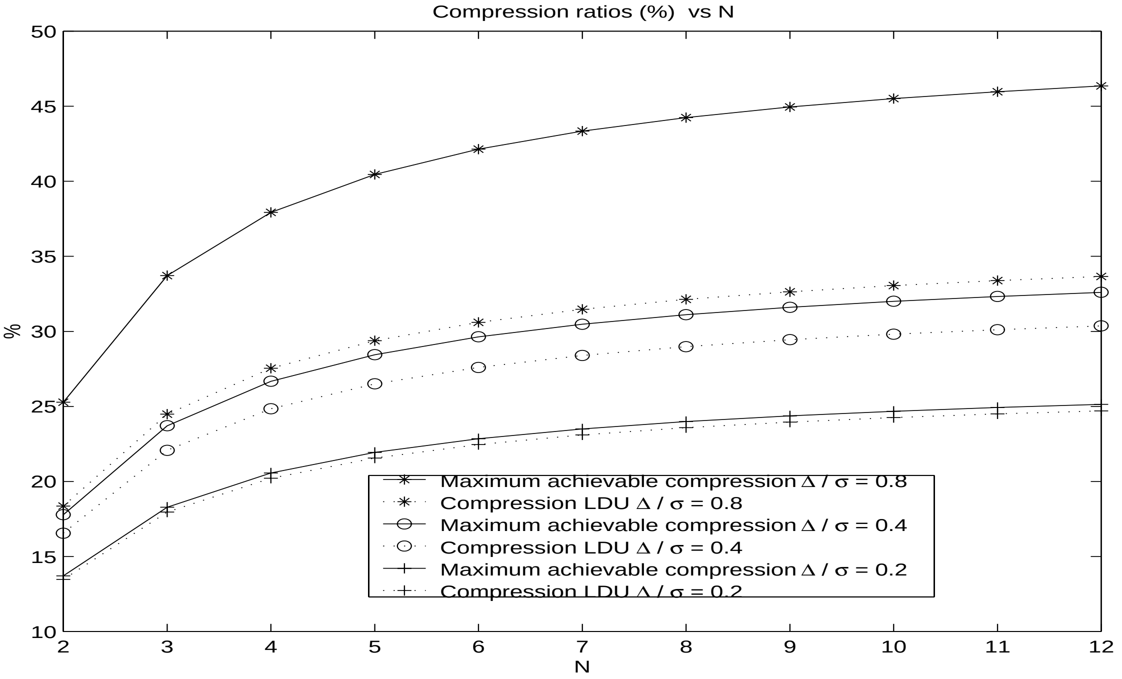


Figure 6.11: Compression ratios for several values of Δ/σ *vs* N for I2I LDU (AR(1) and $\rho = 0.9$).

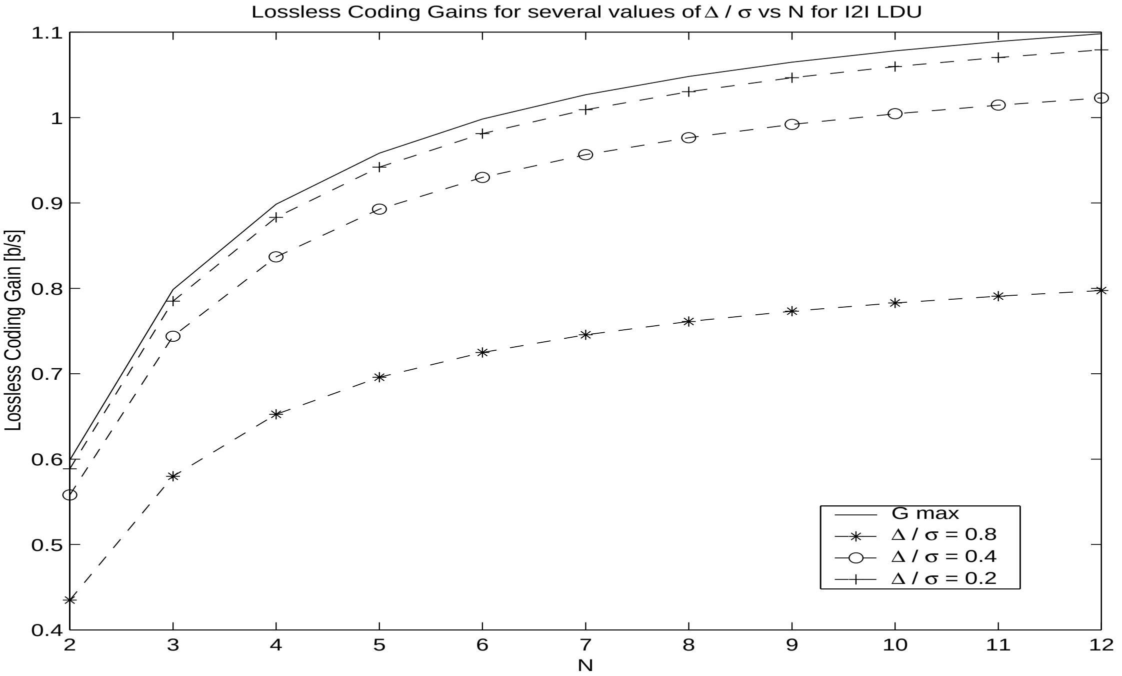


Figure 6.12: Lossless coding gains for several values of Δ/σ *vs* N for I2I LDU (AR(1), $\rho = 0.9$).

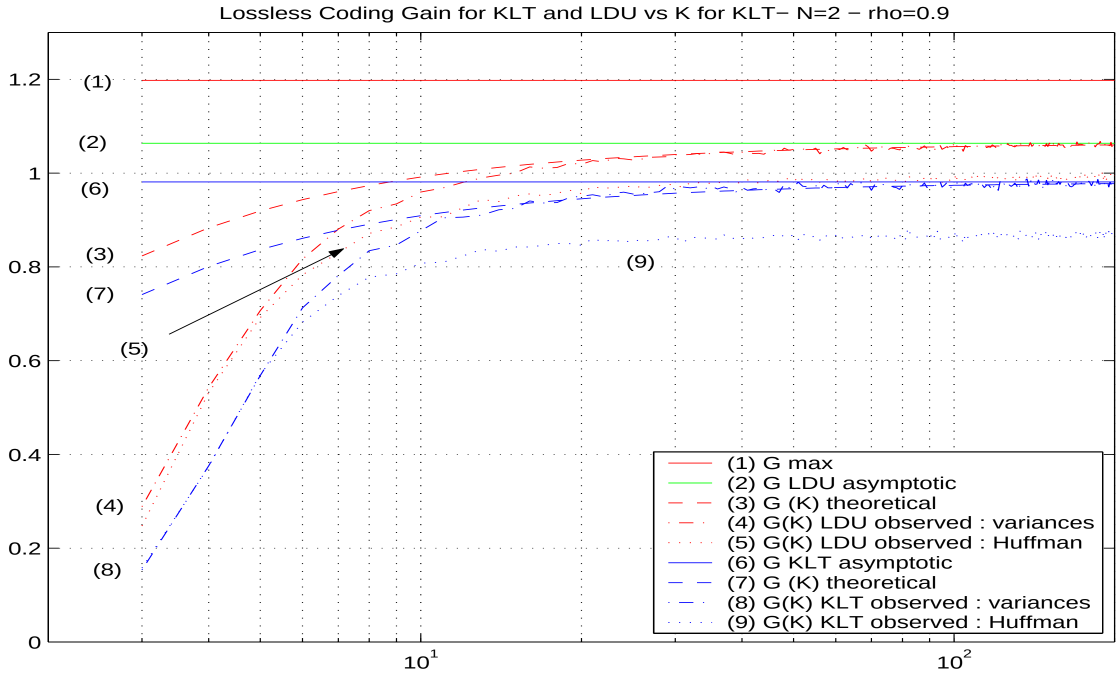


Figure 6.13: Lossless coding gains with estimation noise versus K for $N = 2$. $\frac{\Delta}{\sigma_x} = 0.51$.

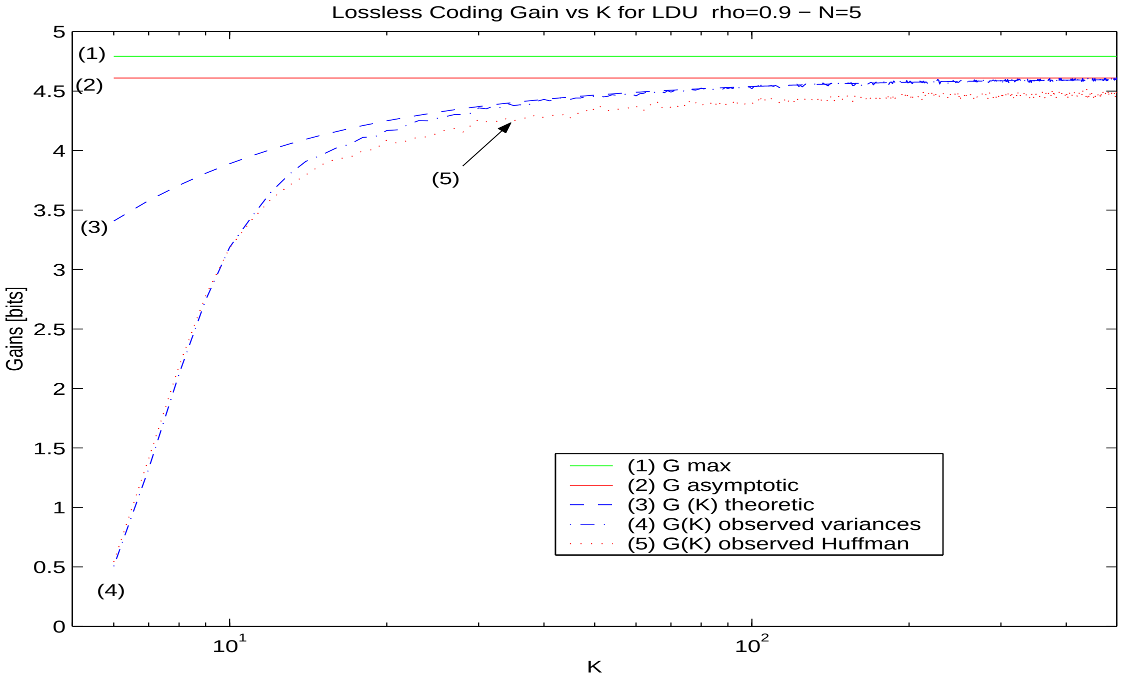


Figure 6.14: Lossless coding gains with estimation noise versus K for N=5. $\Delta = 0.51$.

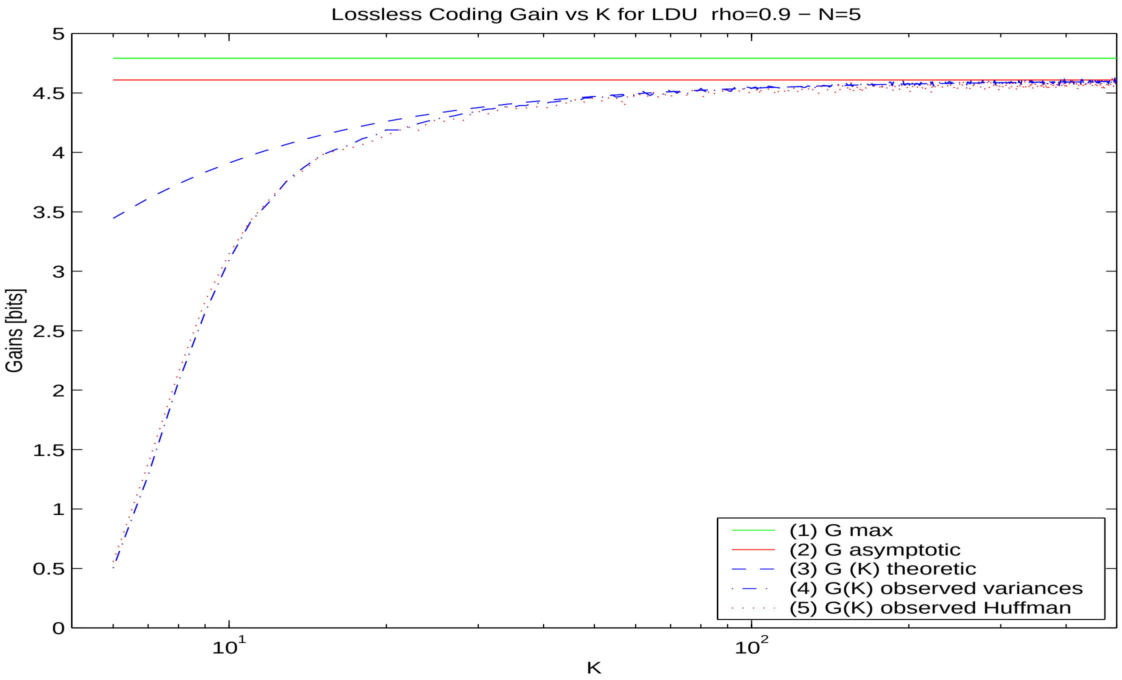


Figure 6.15: Lossless coding gains with estimation noise versus K for N=5. $\Delta = 0.21$.

6.6 Conclusions

For single-stage lossless coding, where the components to be coded are decorrelated by means of integer-to-integer transforms, an upper bound for the lossless coding gain has been described in term of mutual information. The performances of the KLT and the LDU have been described using high resolution approximations, and compared with the maximum achievable coding gain for both fixed and adaptive systems. Various numerical results were then presented. These results indicate that the theoretical analyses regarding the perturbations caused by the lossless constraint, and the estimation noise are fairly accurate as far as the entropies are concerned. The compression performance obtained with Huffman codes are slightly lower than those predicted for both approaches. Moreover, these results show that in any case, the causal transform leads to better compression ratios than its unitary counterpart. Moreover, an interesting side result is that the most coarsely quantized signal should, in the causal case, be placed in first position for the compression to be the most efficient.

Chapter 7

On the Suboptimality of Orthogonal Transforms for Lossless Transform Coding

The analysis of the previous chapter showed that the integer-to-integer implementation of the Karhunen-Loève transform leads to lower compression performance than its causal counterpart. We pursue this analysis in the framework of a multi-stage lossless transform coding scheme, which yields a lossy coded signal, and an error signal. This scheme allows one to choose the respective bitrates of both complementary signals, depending for example on the bandwidth of the transmission link. We show that the causal approach presents several advantages w.r.t. its orthogonal counterparts. For orthogonal transforms, the price paid for the multiresolution approach is a bitrate penalty of 0.25 *bit per sample. This excess bitrate is due to a "gaussianization effect" of the transforms [21]. Firstly, we show under the assumptions of smooth p.d.f.s for the sources, and of high resolution for the lossy coded signal, that the causal approach allows one to code the data (almost) without causing any excess bitrate as compared with a single-stage coder. Secondly, the approach based on the causal transform allows one to easily switch between a single- or a multi-stage compressor. Thirdly, in the framework of interchannel redundancy removal, this approach allows one to easily fix the distortion and rate for both the low resolution and the error signal of each channel, by using different stepsizes in the quantization stage. Any of the channels may, as a particular case, be chosen to be directly losslessly coded. Finally, a side advantage of the causal approach is that entropy coding of the error signal is made very simple since for odd quantization stepsizes, the discrete error sources are uniformly distributed, so that the optimal codewords have the same length, and fixed rate coding is optimal.*

7.1 Introduction

Consider a discrete vectorial source $\underline{x}$ whose samples are $\underline{x}_k$. This source may for example be composed of N scalar signals x_i, in which case $\underline{x}_k = [x_{1,k} \cdots x_{N,k}]^T$, or by the samples of the same scalar source, in which case $\underline{x}_k = [x_k \; x_{k-1} \cdots x_{k-N+1}]^T$. In the framework of a two-stage lossless transform coder each block of signal $\underline{x}_k$ undergoes first a transform, the decorrelated components $\underline{y}_k$ are then quantized by means of uniform scalar quantizers, and further entropy coded, see figure 7.1. The corresponding bitrate will be denoted by $r_{LR}(\underline{y})$.

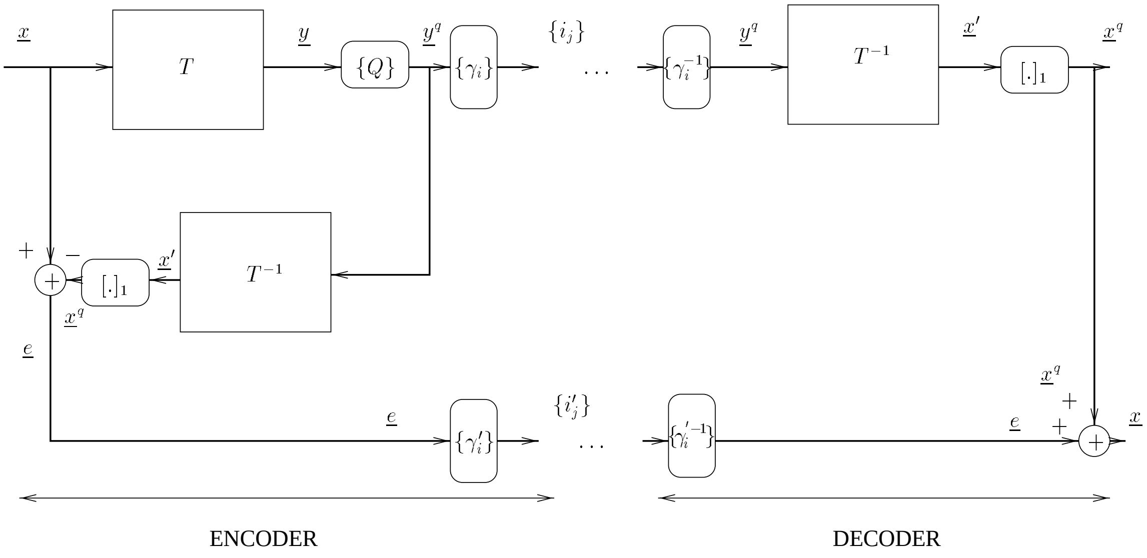


Figure 7.1: Classical two-stage lossless transform coding. $\{Q\}$ denotes uniform scalar quantizers, $\{\gamma_i\}$ and $\{\gamma_i'\}$ scalar entropy coders, and $[.]_1$ rounding operators.

By inverting the transform and taking the integer part of the resulting reconstructed value $\underline{x}'$, the error signal $\underline{e}$ can be generated by substraction : $\underline{e} = \underline{x} - \underline{x}^q$, and further entropy coded. The correspnding bitrate will be denoted by $\overline{r}_{(}\underline{e})$. The decoder generates then $\underline{x}^q$ in the same way, and recovers $\underline{x}$ by $\underline{x} = \underline{x}^q + \underline{e}$. Note that the rounding operations are necessary: since T is a linear transform, $\underline{x}' = T^{-1}\underline{y}^q$ is generally not integer valued.

In this framework, we compare in this chapter the compression performance of orthogonal transforms (*e.g.* DCT, DFT, DST, DHT), as analyzed in [21], to that of the causal transform. A generalization of the two-stage structure to M stages is analyzed for the causal transform in chapter 8.

Let us now denote by $r_{1-shot}(\underline{x})$ the bitrate dedicated to entropy code the source $\underline{x}$ with a single-stage lossless coder. The main question addressed here stands in the following: Is there, in terms of rate, a cost by using any multiresolution approach ? Or in other words, will the overall bitrate $r_{LR}(\underline{y}) + \overline{r}(\underline{e})$ be larger than $r_{1-shot}(\underline{x})$, and if yes, by how much ? The analyses of the next sections will show that the causal transform outperforms orthogonal ones because it avoids the bitrate penalty of 0.25 bit per sample reported in [21], resulting in an optimal system (w.r.t. a single stage coder). Moreover, the causal approach presents several practical important coding features which orthogonal transforms do not share. In the following, the KLT will be used as a benchmark for orthogonal transforms, but as will be underlined, the conclusions of thes analyses can be generalized to other orthogonal transforms.

The rest of the chapter is organized as follows. Section 7.2 states the main assumptions, definitions and notations of this work, and recalls the main characteristics of the causal transform and some results about the "one-shot" compression. Section 7.3 describes the proposed two-stage coding structures and analyzes the statistics of the error signals. Section 7.4 is dedicated to the analysis of the bitrates in the case of Gaussian signals and section 7.5 comments the case of non-Gaussian probability density functions (p.d.f.s). Section 7.6 considers the particular case where lossless transform coding is used to remove intrachannel redundancies, and the last section presents some numerical results.

7.2 Single-Stage Structure

Consider a vectorial source $\underline{x}$, which is obtained by some discretization (quantization) process from a continuous-amplitude source $\underline{x}^c$ (for notation convenience, the time index k will be omitted). In the rest of this chapter, we assume very high resolution (x is integer valued, and $\sigma^2_{x_i} \gg 1$), smooth p.d.f.s for the r.v.s to be coded, and high resolution quantization of the lossy signal ($\Delta_i \ll \sigma^2_{y_i}$).

The rounded value obtained from x^c_i and denoted by $[x^c_i]_1$ is then defined by

$$[x^c_i]_1 = \text{round}(x^c_i) = n,\ n\ \in \mathbb{Z},\ \text{if}\ \ -\frac{n}{2} \leq x^c_i < \frac{n}{2}. \tag{7.1}$$

Similarly, a uniform quantizer with non unity stepsize Δ associates then to x^c_i a quantized value $[x^c_i]_\Delta$. In the case where $\underline{x}^c$ is a vector, $[\underline{x}^c]_\Delta$ will denote quantization of each component x^c_i.

In order to compute the different rates, we will use the Rényi's relation of differential to discrete entropy [38]:

$$H(x_i) + \log_2 \Delta \to h(x^c_i) \text{ as } \Delta \to 0, \tag{7.2}$$

where H denotes the discrete entropy of the discrete source x_i, obtained by uniform quantization with stepsize Δ from the continuous amplitude source x^c_i with differential entropy h. For vectors, a similar relation can be derived, see [35, 162] [1].

We now recall some results of the previous chapters concerning single-stage compression of a vectorial source $\underline{x}$ by means of integer-to-integer transforms.

7.2.1 Lossless Implementation of the Transforms

In the causal case the vector $\underline{x}$ is decorrelated by means of a lower triangular transform L. The transform vector $\underline{y}$ is $L\underline{x} = \underline{x} - \overline{L}\underline{x}$, where $\overline{L}\underline{x}$ is the reference vector. The components y_i are the prediction errors of

[1] Gish and Pierce gave an outline of the proof in [35]; Csiszár generalized the result in [162].

x_i with respect to the past values of $\underline{x}$, the $\{x_{1:i-1}\}$, and the optimal coefficients $-L_{i,1:i-1}$ are the optimal prediction coefficients. It follows that $R_{\underline{xx}} = L^{-1}R_{\underline{yy}}L^{-T}$, which represents the LDU factorization of $R_{\underline{xx}}$. In the unitary case, $R_{\underline{xx}} = V^{-1}\Lambda V^{-T}$, where Λ is the diagonal matrix of the eigenvalues of $R_{\underline{xx}}$. In both cases, $\det R_{\underline{yy}} = \det R_{\underline{xx}}$, since both tranforms are unimodular.
However, since the resulting components y_i are generally not integer, such a transform cannot be used for lossless coding. A lossless implementation of the LDU transform is depicted in figure 7.2.

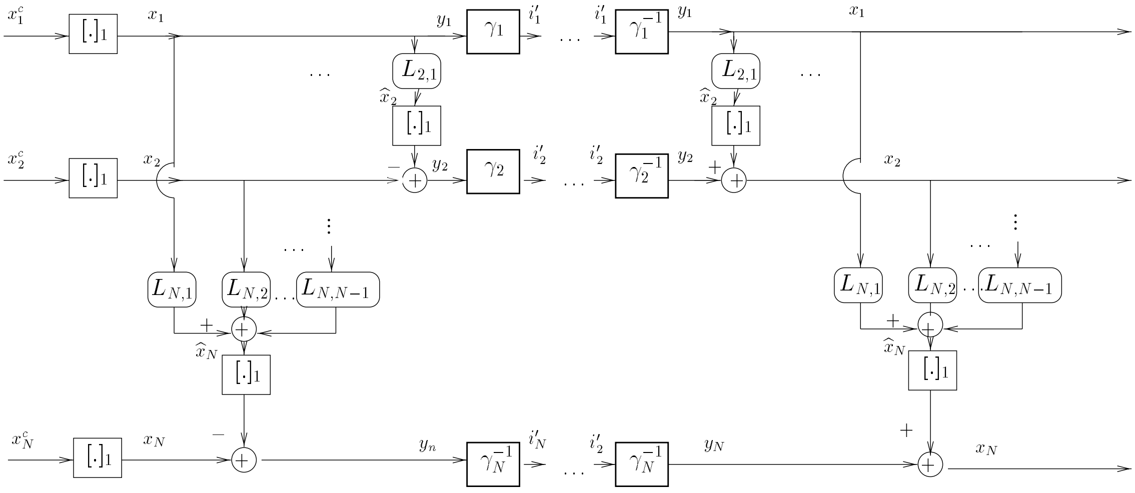


Figure 7.2: Lossless “one-shot” implementation of the LDU Transform.

In this case, the transform signals are obtained by

$$y_{i,k} = x_{i,k} - [\widehat{x}_{i,k}]_1 = x_{i,k} - [\overline{L}_{i,1:i-1}\underline{x}_{1:i-1,k}]_1, \tag{7.3}$$

where $\widehat{x}_{i,k}$ is the estimate of $x_{i,k}$ based on the previous samples of $\underline{x}_k$. The signals y_i are then entropy coded (bitstreams $\{i'_j\}$). At the decoder, each component x_i is losslessly recovered by $x_i = y_i + [\widehat{x}_{i,k}]_1$.

7.2.2 Orthogonal Case

Many lossless implementations of orthogonal transforms have been studied recently, see for example [41, 164, 165]. Concerning the KLT, the integer-to-integer approximation of the optimal linear orthogonal decorrelating transform is based on the factorization of the unimodular matrix into a product of triangular matrices, cascaded with rounding operations ensuring the invertibility of the global transform [41].
Because of its triangular structure, the LDU transform is naturally well suited for factorizations involving lifting steps and roundings. This is not the case for noninteger-valued orthogonal transforms, in which case the number of rounding operations decreases the coding performance. It was shown in [142] that for single-stage coders, the best linear decorrelating orthogonal transform is slightly less efficient than the causal one.

For other transforms such as DCT, DFT, etc, the compression performance will most probably be still worse, since they are square matrices with non-integer coefficients also, and their decorrelation efficiency is less than that of the KLT. In the next section, orthogonal and causal approaches are compared for a two-stage structure.

7.3 Two-Stage Structure

7.3.1 Orthogonal Transforms

As stated in the introduction, the vectorial source $\underline{x}$ can be losslessly coded by means of a two-stage structure, yielding a low resolution version $\underline{x}^q$, and an error signal $\underline{e}$.
In the case of orthogonal transforms (KLT, DCT,...), the coding scheme is represented by figure 5.9. The error signal may be written as $\underline{e} = \underline{x} - \underline{x}^q = [\underline{x} - \underline{x}^{'q}]_1 = [T^{-1}\underline{q}]_1$. Thus, each e_i is a discretized mixture of N random variables (r.v.s), which, as shown by high resolution quantization theory, are uniform if Δ is small in comparison with the variances $\sigma^2_{y_i}$ of the signals y_i, and if their p.d.f.s are smooth. Since the convolution of N uniform r.v.s tends quickly to a Gaussian, the error signals e_i may be approximated as continuous Gaussian r.v.s with variances $\frac{\Delta^2}{12}$, discretized with stepsize unity. The minimum distortion is now obtained by setting $\Delta = 1$, resulting in a distortion of $\frac{\Delta^2+1}{12} = \frac{1}{6}$ on each component. Thus, this scheme does not offer the simple mean of switching from the two-stage to the "one-shot" coder by only setting the quantization stepsizes to 1.
Since the e_i are nearly Gaussian, the probability that an error occurs for a general Δ can be approximated with the error function [21]:

$$P(e_i \neq 0) = P(|e_i| \geq \frac{1}{2}) \approx 1 - erf(\sqrt{\frac{3}{2}}\frac{1}{\Delta}). \tag{7.4}$$

For $\Delta = 1$, this leads to $P(e_i \neq 0) \approx 0.08$, which means that one out of twelve samples should be corrected at the decoder to ensure the losslessness. The question of the rate dedicated to code $\underline{e}$ is examined in the next section.

7.3.2 Causal Transform

The two-stage causal structure may be described by the figure hereafter.

The transform signals are computed by substracting the optimal estimate of x_i based on the past *quantized* samples $x^q_{1:i-1}$, and by quantizing with some stepsize Δ_i the resulting error prediction, which leads to y^q_i. The reason for computing the prediction by means of quantized data is that we are interested in a low resolution signal which can be computed *independently* of the error signals. Thus, only the available x^q_i at the decoder should be used to compute the remaining $x^q_j,\ j > i$. As will be commented in the rate analysis, prediction based on quantized data is slightly less efficient than that based on original data, though

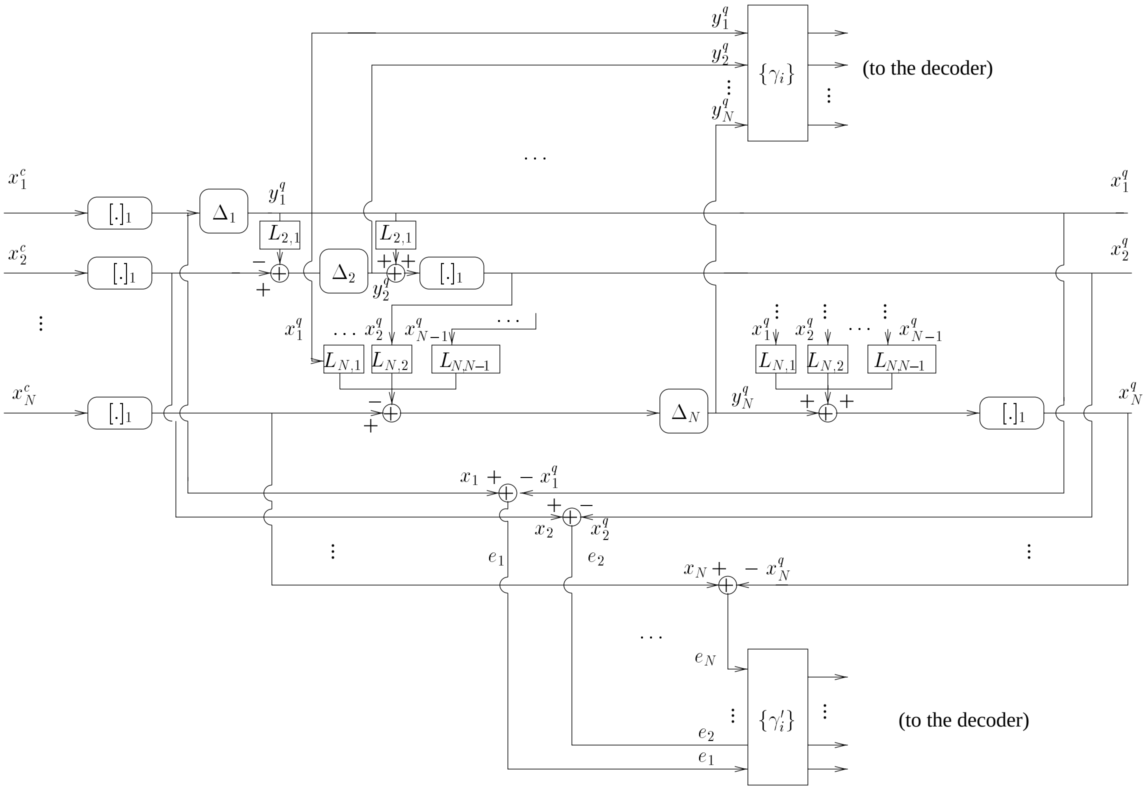


Figure 7.3: Encoder of the two-stage lossless coding Structure in the causal case.

this difference will be shown to be negligible in most of the cases. Each error signal is thus computed by

$$e_i = x_i - x_i^q = x_i - [y_i^q + \overline{L}_{i,1:i-1}\underline{x}^q_{1:i-1}]_1 = [x_i - \overline{L}_{i,1:i-1}\underline{x}^q_{1:i-1} - y_i^q]_1 = [y_i - y_i^q]_1. \qquad (7.5)$$

Thus, the errors e_i are now the discretized versions of the quantization errors in the transform domain. Assuming smooth p.d.f.s and high resolution ($\Delta_i \ll \sigma^2_{y_i}$), three cases should be considered in order to derive the statistics of the errors e_i [2].

Firstly, if $\Delta = 1$, it can be checked that fixing all stepsizes to 1 yields a single-stage lossless coding scheme of figure 7.2. We have noted in the previous section that a similar equivalence is not possible in the case of orthogonal transforms.

If now Δ_i is an odd integer greater than 1, the rounding definition (7.1) yields equally likely errors (with probabilities $p_i^o = \frac{1}{\Delta_i}$), and belonging to $\{-\frac{\Delta_i - 1}{2},\ -\frac{\Delta_i - 1}{2} + 1, ..., \frac{\Delta_i - 1}{2}\}$.

If finally Δ_i is even, all the errors are equally likely except $\pm\frac{\Delta_i}{2}$, which, in virtue of (7.1), and assuming that $P(x_i^q > 0) = P(x_i^q < 0)$, are twice less likely than the other ones (for example, $+\frac{\Delta_i}{2}$ occurs only for positive values of x_i^q). Thus, regarding the probabilities p_i^e of the errors obtained with even Δ, the values

[2] The p.d.f. should not change much within each quantization bin, otherwise the p.d.f.s of the errors may be far from the uniform distribution. Numerical simulations show that this is a reasonable assumption for Gaussian sources.

$0, \pm 1 ... \pm \frac{\Delta_i}{2} - 1$ are equally likely with probabilities $\frac{1}{\Delta_i}$, and $\pm\frac{\Delta_i}{2}$ have probabilities $\frac{1}{2\Delta_i}$. These remarks suggest that the errors will be nonzero with the same probability for even and odd Δ, which is given by

$$P(e_i \neq 0) = P(|e_i| \geq \frac{1}{2}) = 1 - \frac{1}{\Delta_i} \;\; \forall \Delta_i. \tag{7.6}$$

The difference between the cases of even and odd Δ is illustrated in figure 7.4. For the partionning induced by the round off quantizers with $\Delta = 6$, the errors ± 3 are twice less likely than $0, \pm 1, \pm 2$. For $\Delta = 5$, all the cells are equivalent, which makes the errors equally likely.

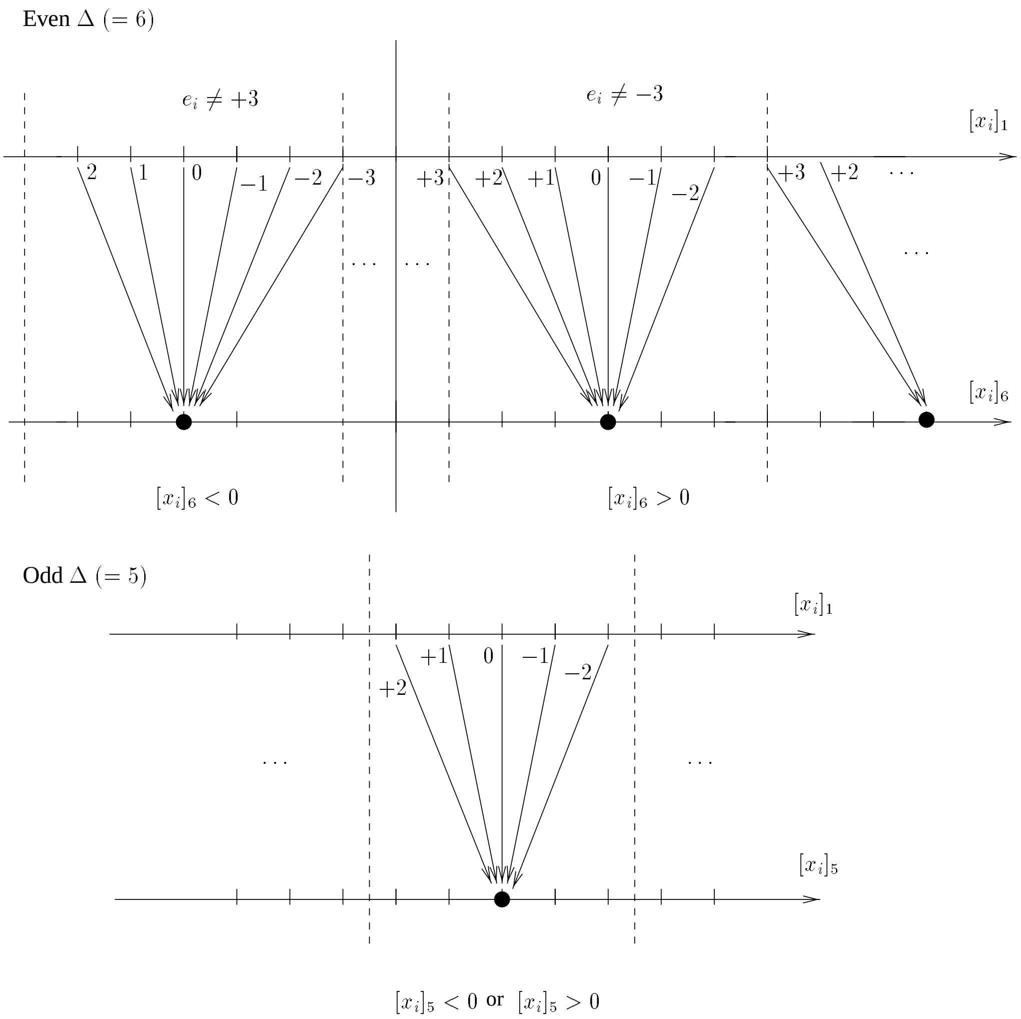


Figure 7.4: Probability of errors induced by the rounding operator (7.1), for even and odd Δ.

Figure 7.5-a) plots the observed and theoretic probabilities of error in the orthogonal case and in the causal case as given by (7.4) and(7.6) (for these simulations, all the quantization stepsizes are equal, see details in section 7.7).

As a conclusion, the causal transform allows one, on the one hand, to switch easily between either a single, or a two-stage structure, by simply fixing the stepsizes to 1. Moreover, the stepsizes Δ_i may in general be different, allowing one to choose a possibly different rate-distortion trade-off for each signal x_i^q. Also, any channel x_i can be chosen in the causal case to be directly losslessly coded, by setting the corresponding Δ_i to 1. On the other hand, the KLT does not benefit from these advantages because of the mixing effect of the quantization errors in the signal domain. As shown in figure 7.5, the probability that an error occurs is higher in the causal case than in the orthogonal case as soon as $\Delta > 1$. The next section will show however that this does not preclude that the rate associated to the error signal is higher in the causal than in the orthogonal case.

7.4 Analysis of the Rates

In this section we assume jointly Gaussian signals for which closed form expression for the rates can be obtained; the case of non-Gaussian p.d.f.s will be discussed in section 7.6. Moreover, we assume for simplicity that all the quantization stepsizes are equal in both causal and unitary cases (though this is not necessary for the LDU transform, as stated in section 7.3).

7.4.1 Low Resolution Versions

For the two transformations, one should compute

$$r_{LR_T} = \frac{1}{N}\sum_{i=1}^{N} H(y_i, T) \approx \frac{1}{N}\sum_{i=1}^{N} h(\sigma_{y_i}^2, T) - \log_2 \Delta, \tag{7.7}$$

where T denotes either the causal or the unitary transform. For both transforms, the transform signals are Gaussian. The variances $\sigma_{y_i}^2$ are in the orthogonal case the eigenvalues λ_i of $R_{\underline{xx}}$, so that

$$r_{LR_V} \approx \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\log_2 2\pi e\lambda_i - \log_2 \Delta\right) \approx \frac{1}{2}\log_2 2\pi e(\det R_{\underline{xx}})^{\frac{1}{N}} - \log_2 \Delta. \tag{7.8}$$

In the causal case, the variances of the transform signals $\sigma_{y_i}^2$ are not exactly the optimal prediction error variances $\sigma_{y_i^0}^2$ of order $i-1$ based on $x_{1:i-1}$, because the prediction is computed by means of quantized samples. One shows that (see result (2.36) with $\sigma^2 = \frac{\Delta^2}{12}$) $\sigma_{y_i}^2 \approx \sigma_{y_i^0}^2 + \frac{\Delta^2}{12}(\overline{LL^T})_{ii}$. As in DPCM, the prediction error variances are increased due to a quantization noise feedback.Using (2.68), we obtain

$$\begin{aligned} r_{LR_L} &= \tfrac{1}{N}\sum_{i=1}^{N} H(y_i, L) \\ &\approx \tfrac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\log_2 2\pi e\sigma_{y_i}^2 - \log_2 \Delta\right) \\ &\approx \tfrac{1}{2}\log_2 2\pi e\left(\prod_{i=1}^{N}\sigma_{y_i^0}^2\right)^{\frac{1}{N}}\left[1 + \tfrac{\Delta^2}{12N}\sum_{i=1}^{N}(\overline{LL^T})_{ii}\right] - \log_2 \Delta \\ &\approx \tfrac{1}{2}\log_2 2\pi e(\det R_{\underline{xx}})^{\frac{1}{N}} + \tfrac{\Delta^2}{24N\ln 2}\sum_{i=1}^{N}\left(\frac{1}{\lambda_i} - \frac{1}{\sigma_{y_i^0}^2}\right) - \log_2 \Delta. \end{aligned} \tag{7.9}$$

Thus, for the same distortion $\frac{\Delta^2+1}{12}$ on each component x_i^q, the bitrate required to entropy code the low resolution version obtained by means of the causal transform should require an excess bitrate in comparison with the KLT. Simulations in section 7.7 show however that this excess bitrate is negligible in many practical coding situations.

7.4.2 Error Signals

Regarding now the rate $\overline{r}_T$ dedicated to the error signals, one can compute the discrete entropies of the signals e_i by using the error analysis of section 7.3.

In the unitary case, each e_i can be seen as a discretized Gaussian r.v. with variance $\frac{\Delta^2}{12}$. Thus, the bitrate $\overline{r}_V = \frac{1}{N}\sum_{i=1}^{N} H(e_i, V)$ can be written as [21]

$$\overline{r}_V \approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\log_2 2\pi e\frac{\Delta^2}{12} = \log_2\Delta + \underbrace{\frac{1}{2}\log_2\frac{\pi e}{6}}_{\approx 0.25\ bit}, \tag{7.10}$$

We find in (7.10) the well known difference between Gaussian and uniform entropies [29] of ≈ 0.25 bit.[3].

In the causal case we obtain, depending on the parity of Δ

$$\begin{aligned} \overline{r}_{L,even} &= -\sum_{i=1}^{N} p_i^e \log_2 p_i^e \approx -(\Delta-1)\frac{1}{\Delta}\log_2\frac{1}{\Delta} - 2\left(\frac{1}{2\Delta}\log_2\frac{1}{2\Delta}\right) \approx \log_2\Delta + \frac{1}{\Delta}\ , \\ \overline{r}_{L,odd} &= -\sum_{i=1}^{N} p_i^o \log_2 p_i^o \approx -\Delta\left(\frac{1}{\Delta}\log_2\frac{1}{\Delta}\right) \approx \log_2\Delta. \end{aligned} \tag{7.11}$$

Comparing (7.10) and (7.11), the approximately 0.25 bit/sample excess rate of orthogonal transforms w.r.t. single-stage lossless coding vanishes in the causal case. Moreover, in the case of odd Δ, the error are uniformly distributed, which means that no compression is required for the bitrate to reach the entropy of the sources e_i, and the optimal coding procedure is simply consists in transmitting the binary representation of the values e_i.

7.5 Intrachannel Redundancy Removal

The coding schemes presented in figures 7.2 and 7.3 can indeed be used to remove intrachannel redundancies, in which case frequential expression can be obtained. In this case, each data block is $\underline{x}_k = [x_k\ x_{k-1}\cdots x_{k-N+1}]^T$. Again, we assume a Gaussian p.d.f. and equal quantization stepsize Δ for the quantizers $\{Q\}$. By letting the block length grow to infinity, and using the asymptotic distribution of Toeplitz matrices [166],

$$\lim_{k\to\infty}\det(R_{\underline{xx}})^{\frac{1}{N}} = \lim_{k\to\infty} e^{\frac{1}{N}\log\prod_{i=1}^{N}\lambda_i} = e^{\int_{-\frac{1}{2}}^{\frac{1}{2}}\ln S_{xx}(f)df}, \tag{7.12}$$

[3] This (often called “quarter bit”) result was first reported by Koshelev in [29], rediscovered by numerical simulations by Goblick and Holsinger [30] and derived analytically by Gish and Pierce [35].

where $S_{xx}(f)$ denotes the power spectral density of x, we get for the bitrates of the low resolution signals

$$\begin{array}{rcl} r_{LR_V} & \approx & \frac{1}{2}\log_2 2\pi e\; e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln S_{xx}(f)df} - \log_2 \Delta, \\ r_{LR_L} & \approx & r_{LR_V} + \frac{\Delta^2}{24\ln 2}\left[\int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}^{-1}(f)df - e^{-\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln S_{xx}(f)df}\right], \end{array} \tag{7.13}$$

The bitrates corresponding to the error signals (7.10) and (7.11) remain unchanged.

7.6 Case of Non-Gaussian p.d.f.s

Regarding the low resolution signals, non-Gaussian p.d.f.s of the x_i may lead to non-Gaussian p.d.f.s for the y_i[4]. Since the relation of the differential entropies to the variances of the transform signals will be different from that of Gaussian r.v.s, the rates r_{LR_V} and r_{LR_L} will differ from equation (7.8) and (7.9). However, since a Gaussian r.v. maximizes the differential entropy for a given variance, one may expect that the actual rates will be lower than those of equation (7.8) and (7.9), obtained in the Gaussian case.

As for the error signals, the analyses of the previous sections are still valid under the same assumptions of smooth p.d.f.s and high resolution. The quantization errors in the transform domain are still uniform, leading, in the signal domain, to nearly Gaussian errors in the orthogonal case, and to nearly uniformly distributed errors in the causal case. Thus the causal approach avoids the 0.25 bit suboptimality of the orthogonal transforms regardless of the p.d.f.s of the sources.

7.7 Numerical Results

For the simulations, we generated 10^5 real Gaussian i.i.d. vectors with covariance matrix $R_{\underline{x}\underline{x}} = HR_{AR1}H^T$. R_{AR1} is the covariance matrix of an AR(1) process with $\rho = 0.9$ and variance 10^5. H is a diagonal matrix whose ith entry is $(N-i+1)^{1/3}$, $N = 3$. The data are rounded with even or odd Δ. A "one-shot" approach requires ≈ 9.8 b/s to losslessly code these data.

Figure 7.5-b) compares the theoretic (expression (7.8) for the KLT, and (7.9) for the LDU) and observed entropies for the low resolution signals. Note that the excess rate in the causal case (*cf* equation (7.9)) is negligible as long as r_{LR_L} is greater than roughly 3 bits/sample, which is one third of the overall bitrate. The first set of figures deals with odd Δ. Figure 7.6-a) compares the theoretic (expressions (7.10) and (7.11)) and observed entropies for the error signals. Figure 7.6-b) compares the theoretic and observed overall rates for the two-stage coders in both approaches, showing that the best orthogonal approach is nearly 0.25 bit suboptimal w.r.t. its causal counterpart in most cases.

In the case of even Δ, similar results are obtained in figures 7.7 and 7.8. Note the excess rate term ($\frac{1}{\Delta}$) which appears in $\overline{r}_{L,even}$ of equation (7.11) for low values of Δ. As shown in figures 7.8a and 7.8b, choosing the

[4] *e.g.*, if $\underline{x} = A\underline{z}$ is a rotated version of some $\underline{z} = [z_1 \cdots z_N]^T$, where z_i are independent and uniformly distributed, then the KLT V will be $V = A^{-1}$, and $\underline{y}$ will equal $\underline{z}$.

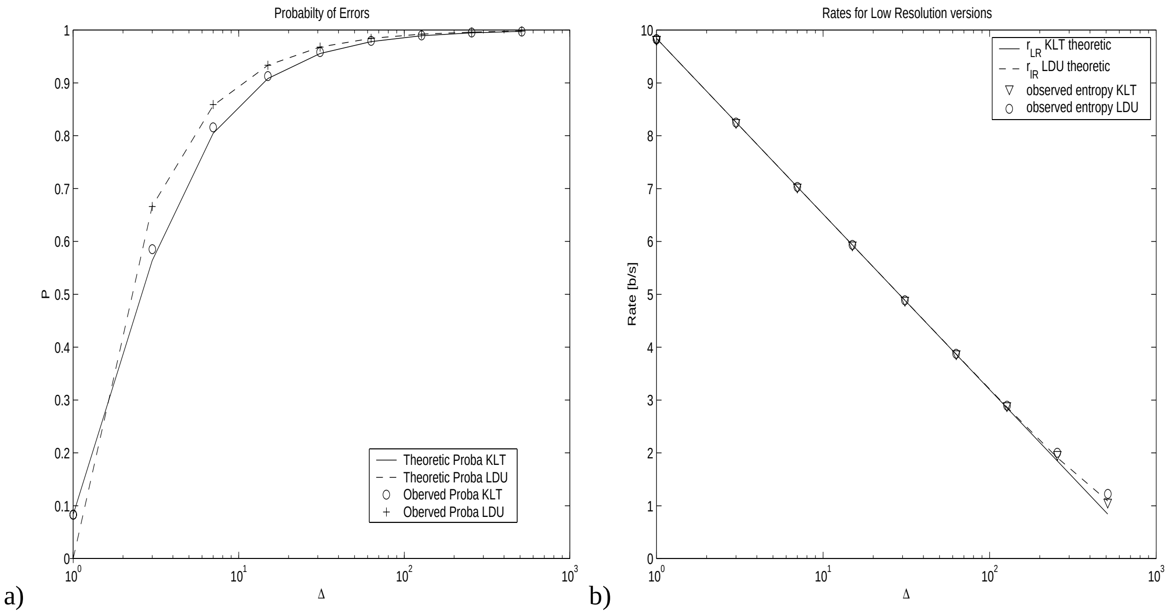


Figure 7.5: Case of odd Δ: a) Error probability and b) Entropies of low resolution versions.

causal approach is preferable if one desires to transmit a lossy signal whose rate is less than approximately 7.8 bits per sample.

Finally, these results are confirmed by figures 7.9, 7.10 and 7.11, where the rates are the actual rates obtained by Huffman coding the different signals.

7.8 Conclusions

The causal LDU transform has been shown to present several advantages over orthogonal transforms in the framework of multi-stage lossless transform coding. Firstly, under the assumption of smooth p.d.f.s for the sources, and of high resolution for the lossy coded signal, the causal approach allows one to code the data (almost, that is, neglecting the noise-feedback term in (7.9) and (7.13)) without causing any excess bitrate as compared with a single-stage coder. Secondly, the approach based on the causal transform allows one to easily switch between the single-stage compressor described in chapter 6 or a multi-stage lossless coder. Thirdly, in the framework of interchannel redundancy removal, this approach allows one to easily fix the distortion and rate for both the low resolution and the error signal of each channel, by using different stepsizes in the quantization stage. Any of the channels may, as a particular case, be chosen to be directly losslessly coded. Finally, a side advantage of the causal approach is that entropy coding of the error signal is made very simple, since for odd quantization stepsizes, the discrete error sources are uniformly distributed, so that the optimal codewords have the same length, and fixed rate coding is optimal.

Indeed, better compression performance may be obtained by removing intra- in addition to inter-channel redundancies if the vectorial source $\underline{x}$ presents memory. The next chapter presents the extention of the previous results in this case.

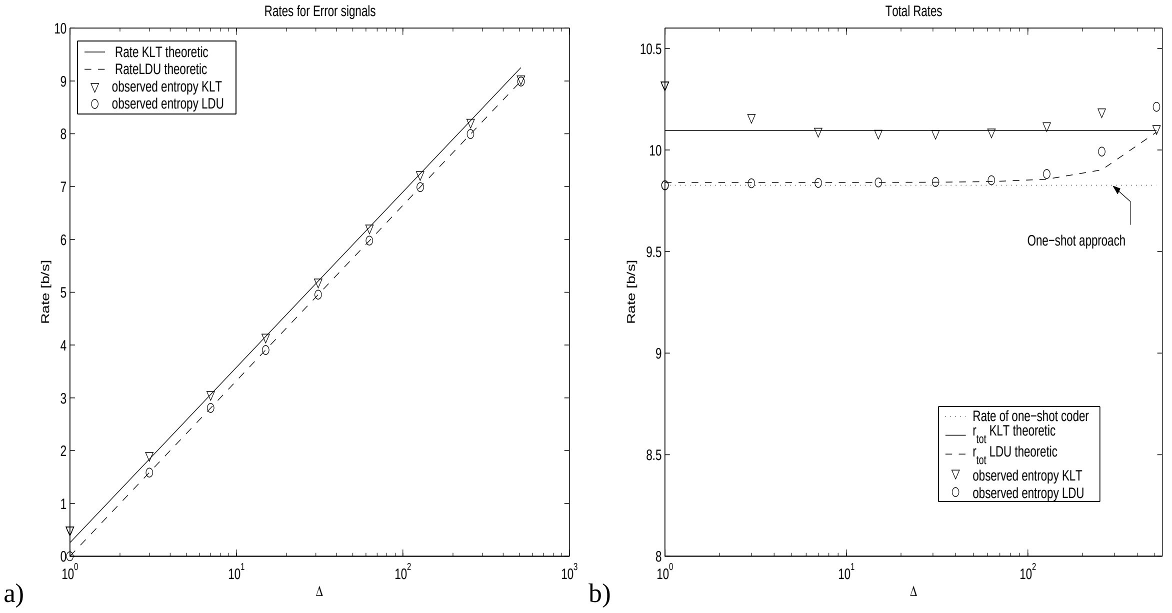


Figure 7.6: Case of odd Δ: a) Entropies for error signals and b) Overall entropies.

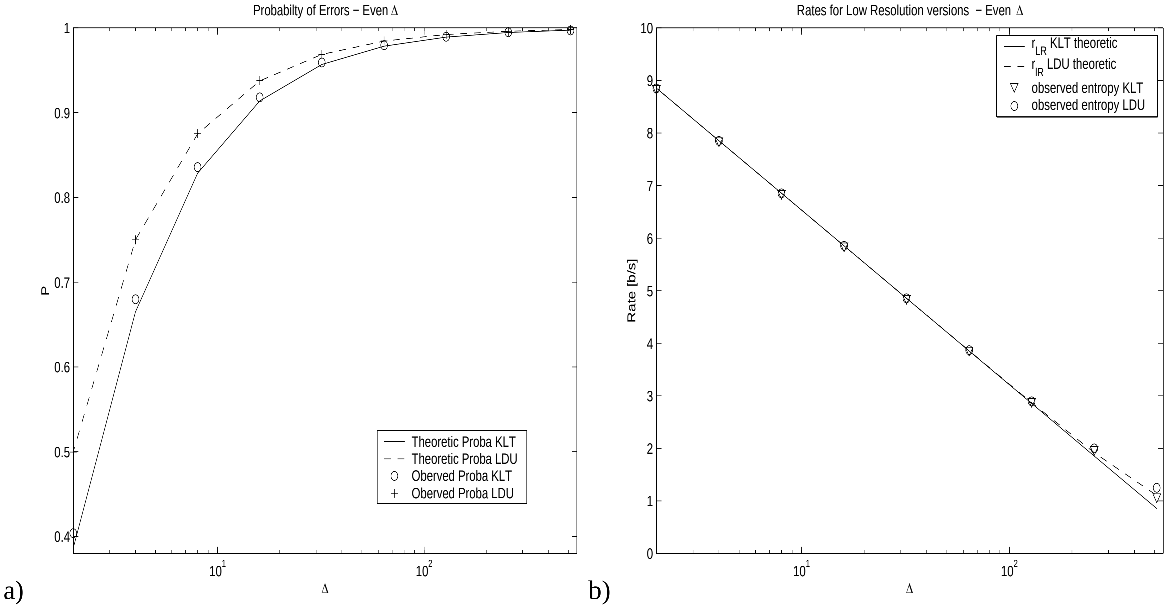


Figure 7.7: Case of even Δ: a) Error probability and b) Entropies for low resolution versions.

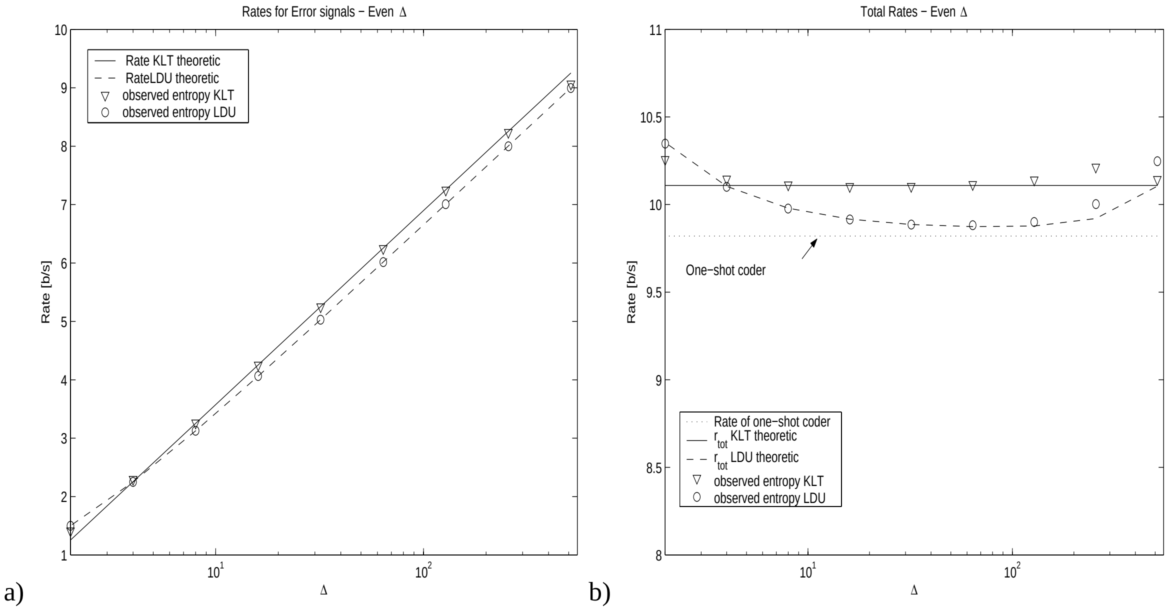


Figure 7.8: Case of even Δ: a) Entropies of the error signals and b) Overall entropies.

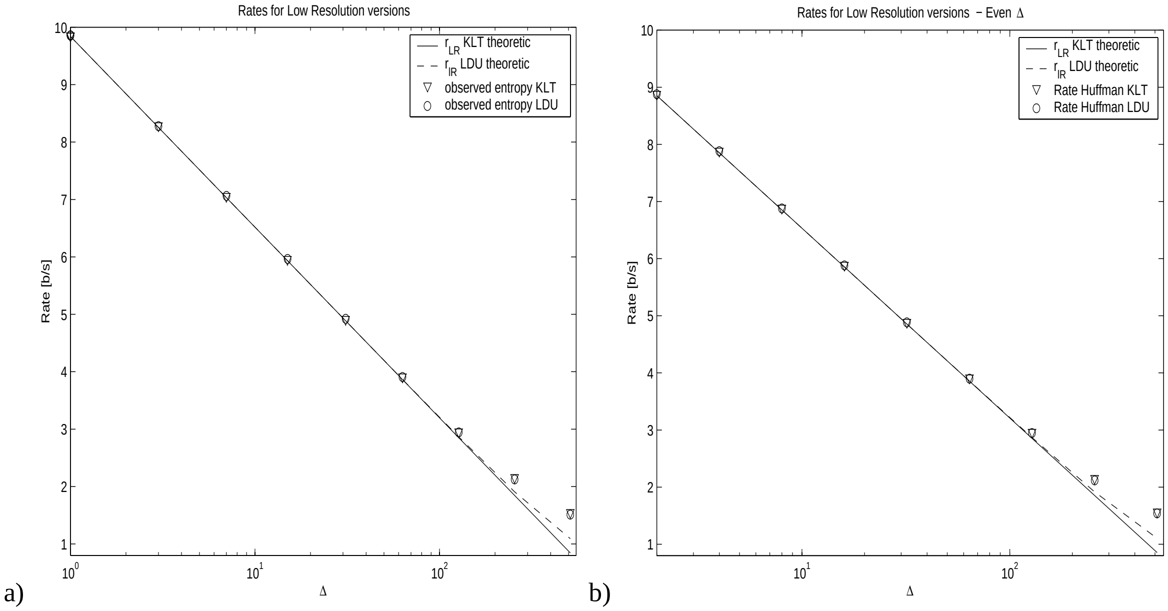


Figure 7.9: Rates obtained by Huffman coding for low resolution versions: a) Odd Δ and b) Even Δ.

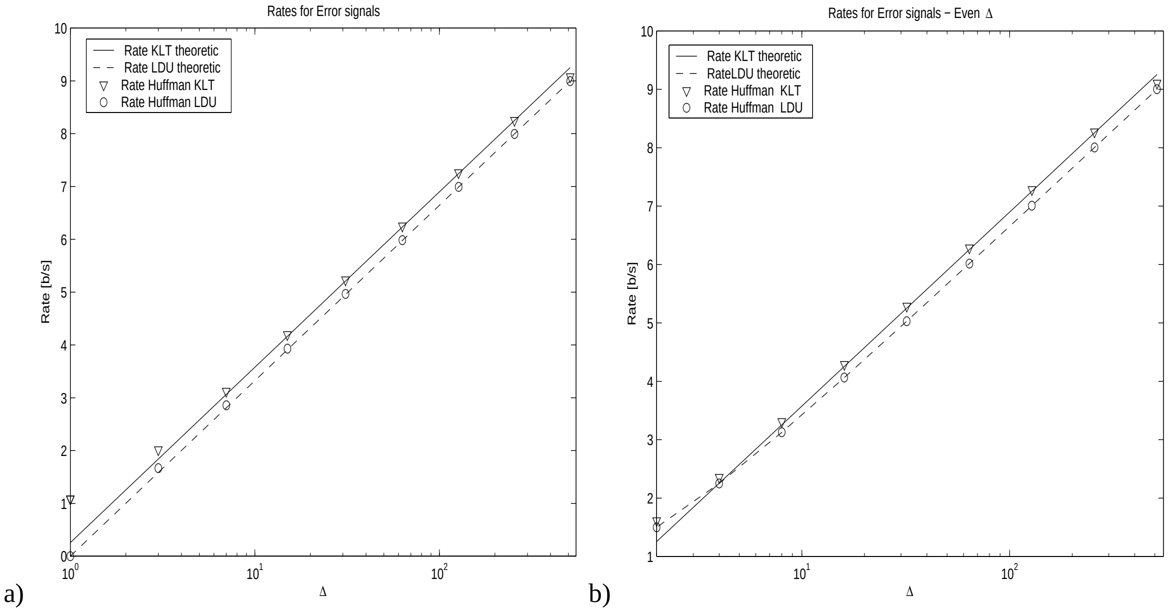


Figure 7.10: Rates obtained by Huffman coding for error signals: a) Odd Δ and b) Even Δ.

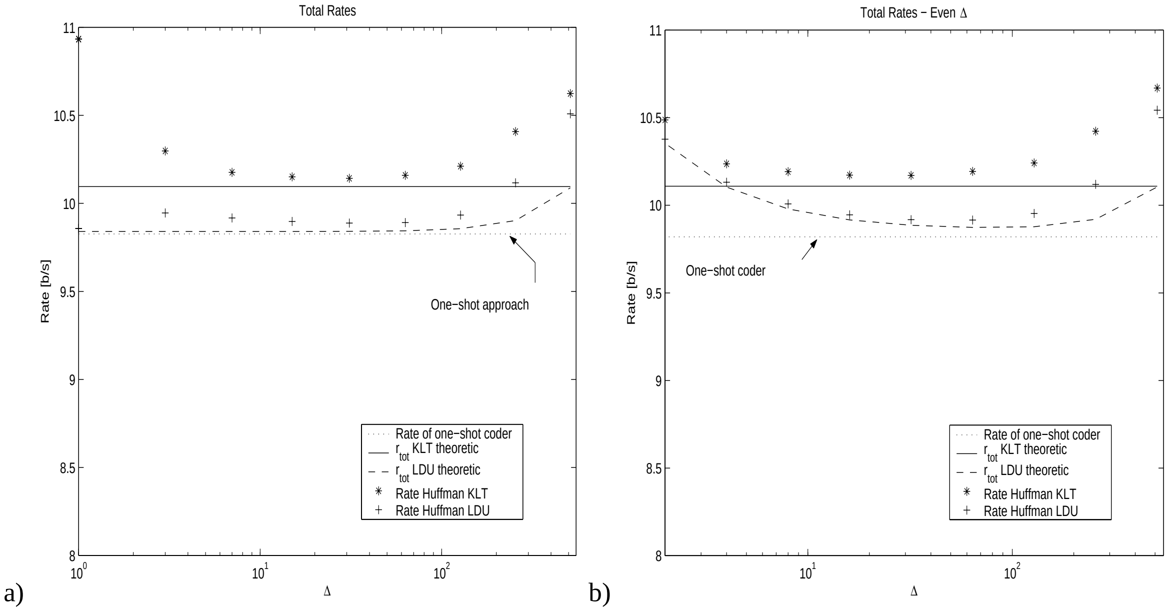


Figure 7.11: Overall Rates obtained by Huffman coding: a) Odd Δ and b) Even Δ.

Chapter 8

Multistage Integer-to-Integer MIMO Prediction

This chapter investigates lossless coding procedures based on the "generalized MIMO prediction" as analyzed in chapter 5, and on the single- and multi-stage lossless coders of chapters 6 and 7. The considered coding schemes are applied to discrete vectorial sources with memory. In this case, both intra- and inter-channel redundancies are removed by lossless prediction. The resulting signals are scalar entropy coded. For Gaussian sources discretized with uniform scalar quantizers Q_i, we establish first the expression of the maximal bitrate reduction as achievable by any lossless coding technique. This bound corresponds to the performance of optimal vector entropy codes. We compare then the performance of the described integer-to-integer MIMO prediction lossless coding schemes to this bound. Theses schemes are suboptimal because of the lossless constraint imposed to the transformations, which vanishes in the limit of small distortions introduced by the quantizers Q_i. The proposed coders may be used either as compressors, or as a scalable lossless coder. In the latter case, a multistage version of the lossless coder based on triangular MIMO predictor is proposed. (A)DPCM lossless prediction loops are introduced which allow one to transmit the data by means of substreams, which represent different "resolution" levels. This multiresolution approach is slightly suboptimal in comparison with a single-stage compression approach because of the noise feedback created in the (A)DPCM loops. We propose a strategy to fix the stepsizes of the quantizers of these loops so that the delivered rates approach some predetermined target rates.

8.1 Introduction

8.1.1 Lossless Coding

Let us consider a continuous-amplitude Gaussian vectorial source $\underline{x}^c$. In a first step, this source is quantized, resulting in a source $\underline{x}$[1]. As $\underline{x}^c$, the source $\underline{x}$ may present both temporal and spatial dependencies.
Once some rate-distortion trade-off has been chosen, the distortion is fixed. By the noiseless coding theorem of Shannon, the minimum bitrate r_0 required to code the discrete-amplitude source $\underline{x}$ corresponds to its entropy rate. The aim of lossless coding is to design a coding procedure whose actual bitrate will be as small as possible, and, if possible, will reach r_0. Indeed, it is known that entropy coders which assign adequate codewords to blocks of samples $\underline{x}_k$, according to the joint probability of these vectors, can reach r_0. The complexity of these vector entropy coders may, however, be prohibitive. Thus, an interesting question is that of designing a coding procedure which is performant in terms of rates, though maintaining a reasonnable complexity, by using scalar entropy coders. This problem was investigated in chapter 6, where we analyzed the performances of lossless transforms (based on the KLT and on the LDU), followed by scalar entropy coders. The first topic of the present chapter is to analyze the performance of similar coding schemes where the transform $T(z)$ corresponds this time to the decorrelation approaches of the MIMO prediction framework. The corresponding single-stage lossless structure is recalled in figure 8.1.

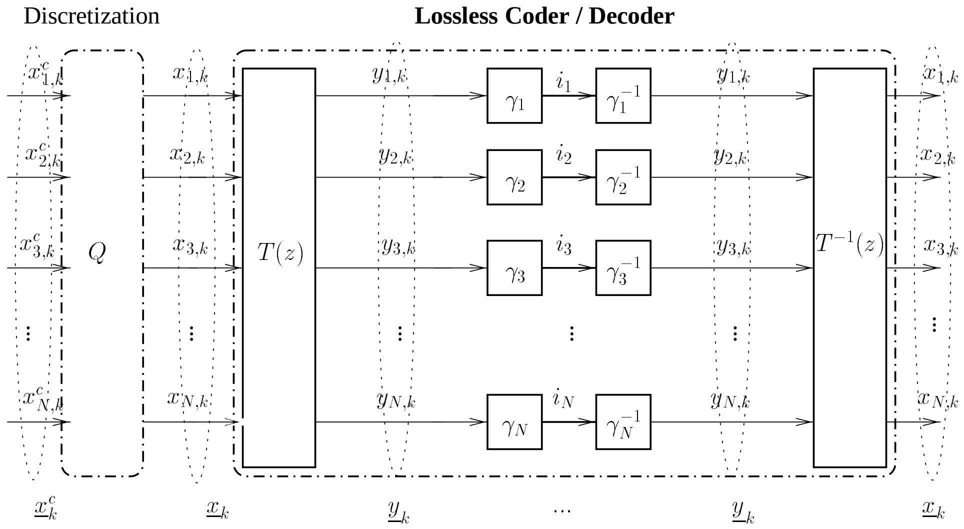


Figure 8.1: Lossless coding scheme considered in this chapter.

Assume in a first scenario that the components $x_{i,k}$ of the vectors $\underline{x}_k$ in figure 8.1 are directly scalar entropy coded (entropy coders γ_i), resulting in a bitrate $r_{scal}(\underline{x})$. Assume in a second scenario that a reversible transformation $T(z)$, aimed of removing intra- and inter-channel dependencies, is applied to

[1] The subscript ^{q} will be dropped for discrete sources x_i and y_i, and will be later dedicated for the DPCM quantized signals in the multistage structure of 8.5.

$\underline{x}_k$ before scalar entropy coding, resulting in a bitrate $r_{scal}(\underline{y})$. As in chapter 6, one may define for this transform a lossless coding gain expressed in bits per sample as

$$G_{T(z)} = r_{scal}(\underline{x}) - r_{scal}(\underline{y}). \tag{8.1}$$

As in chapter 6 also, one may expect that $G_{T(z)}$ is upper bounded by some $G_{Max} = r_{scal}(\underline{x}) - r_0$. The general expression of G_{Max} for Gaussian sources with memory will be derived in section 8.2. This gain will be compared to the lossless coding gain of integer-to-integer implementations of totally decorrelating MIMO predictors in section 8.3.

After the analysis of these "one-shot" lossless structures, the fourth part will turn to two-stages multichannel prediction structures, in the spirit of those described in chapter 7. The bitrates for both the low resolution and the error signals will be first evaluated; the overall bitrate will then be compared to that of the corresponding "one-shot" lossless coders. The two-stage structure will be extended to M stages in section 8.5. Finally, some numerical results will be presented in section 8.6.

8.2 Entropy Rates and Maximum Lossless Coding Gain

The aim of this section is to establish the maximum bitrate reduction, or lossless coding gain, as achievable by any lossless coding method over $r_{scal}(\underline{x})$. We first derive the minimal rate $r_0(\underline{x})$ required to represent the discrete-amplitude, N-dimensional source $\underline{x}$, obtained from $\underline{x}^c$ by some discretization process (figure 8.1). We will then express the bitrate $r_{scal}(\underline{x})$.

By the noiseless coding theorem of Shannon, the minimal bitrate required to represent the source $\underline{x}$ is

$$\min\{r\} = r_0(\underline{x}) + \epsilon \ \text{ bits per sample}, \tag{8.2}$$

where $r_0(\underline{x})$ denotes the entropy rate, and ϵ is a positive value which can be made arbitrarily close to zero by means of optimal vector entropy coders. We assume that $\underline{x}$ is an uniformly quantized version of $\underline{x}^c$ with stepsizes Δ_i. Let the samples of $\underline{x}$ be collected in a vector $\underline{X}_k = [\underline{x}_1 ... \underline{x}_k]^T$ and denote by $\underline{X}_k^c$ the corresponding vector of samples for $\underline{x}^c$. The entropy rate $r_0(\underline{x})$ is defined by the limit

$$r_0(\underline{x}) = \lim_{k\to\infty} \frac{1}{Nk} H(\underline{X}_k). \tag{8.3}$$

Now, for any continuous-amplitude source x_i^c uniformly quantized with stepsize Δ_i, the differential entropy $h(x_i^c)$ can be related to the discrete entropy $H(x_i)$ by the Rényi's relation [38]

$$H(x_i) + \log_2 \Delta_i \to h(x_i^c) \ \ as \ \ \Delta_i \to 0. \tag{8.4}$$

This result can be extended to the Nk-vector $\underline{X}_k$ [35, 162], leading to

$$r_0(\underline{x}) \approx \lim_{k\to\infty} \frac{1}{Nk} h(\underline{X}_k^c) + \frac{1}{N}\sum_{i=1}^{N} \log_2 \Delta_i. \tag{8.5}$$

Expressing the differential entropy of the multivariate normal distribution (see, e.g., [3] pp. 230) we obtain

$$r_0(\underline{x}) \approx \lim_{k\to\infty} \frac{1}{2}\log_2 2\pi e \left(\det R_{\underline{X}_k^c \underline{X}_k^c}\right)^{\frac{1}{Nk}} + \frac{1}{N}\sum_{i=1}^{N}\log_2 \Delta_i, \tag{8.6}$$

Using the result (5.7), the minimum bitrate required to code the source $\underline{x}$ can be expressed as

$$r_0(\underline{x}) \approx \frac{1}{2}\log_2 2\pi e \left(e^{\frac{1}{N}\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\det S_{\underline{x}^c\underline{x}^c}(f)} df \right) + \frac{1}{N}\sum_{i=1}^{N}\log_2 \Delta_i. \tag{8.7}$$

where $S_{\underline{x}^c\underline{x}^c}(f)$ is the power spectral density of the vectorial process $\underline{x}^c$.

As mentioned previously, this bitrate can be achieved by optimal vector entropy coding. If now we use scalar entropy coders to code the x_i, the bitrate is that of expression (6.7)

$$r_{scal}(\underline{x}) \approx \frac{1}{2}\log_2 2\pi e \big(\det(\,\text{diag}\,\{R_{\underline{x}^c\underline{x}^c}\})\big)^{\frac{1}{N}} - \frac{1}{N}\sum_{i=1}^{N}\log_2 \Delta_i. \tag{8.8}$$

Finally, the maximum lossless coding gain corresponding to the bitrate reduction achieved by vector over scalar entropy coders is

$$\begin{aligned} G_{Max} &= r_{scal}(\underline{x}) - r_0(\underline{x}) \\ &\approx \tfrac{1}{2N}\log_2 \frac{\det\ \text{diag}\,\{R_{\underline{x}^c\underline{x}^c}\}}{e^{\int_{-1/2}^{1/2}\ln\det S_{\underline{x}^c\underline{x}^c}(f)df}} \\ &\approx \tfrac{1}{2}\log_2 G_L^{(0)}. \end{aligned} \tag{8.9}$$

where $G_L^{(0)}$ is the optimal coding gain (5.9) obtained in chapter 5, corresponding to an optimal decorrelation of the source *before* the quantization stage. This expression generalizes (6.7), which links similarly the coding gains of the classical and lossless transform coding frameworks. Note that for uniform quantization, G_{Max} does not depend on the stepsizes (assuming they are sufficiently small), but on the spatial and temporal dependencies of the continuous amplitude sources x_i^c only. Also, (6.7) is indeed a special case of (8.9), since for memoryless sources, $S_{\underline{x}^c\underline{x}^c}(f)$ becomes $R_{\underline{x}^c\underline{x}^c}$, and $e^{\int_{-1/2}^{1/2}\ln\det S_{\underline{x}^c\underline{x}^c}(f)df}$ reduces to $\det R_{\underline{x}^c\underline{x}^c}$. The next section investigates the coding gain of actual transforms based on MIMO prediction, followed by scalar entropy coders.

8.3 “One-Shot” Integer-to-Integer Multichannel Prediction

We first present the general structures corresponding to “one-shot” integer-to-integer multichannel prediction. The corresponding coding gain will then be computed, and compared to G_{Max} of (8.9).

8.3.1 Triangular MIMO prediction

The causal decorrelation approaches presented in chapter 6 are easily adapted to lossless coding by introducing round off quantizers, similarly to those presented in figure 6.3. The choice of a particular structure of the generalized MIMO prediction framework depends then on the degrees of non causality which are attributed to the intersignals filters. As in the lossy coding case, particular structures are the triangular, and the classical lossless MIMO predictors.

The application of the triangular MIMO predictor to lossless coding is depicted in figure (8.2) for a two dimensional vector source.

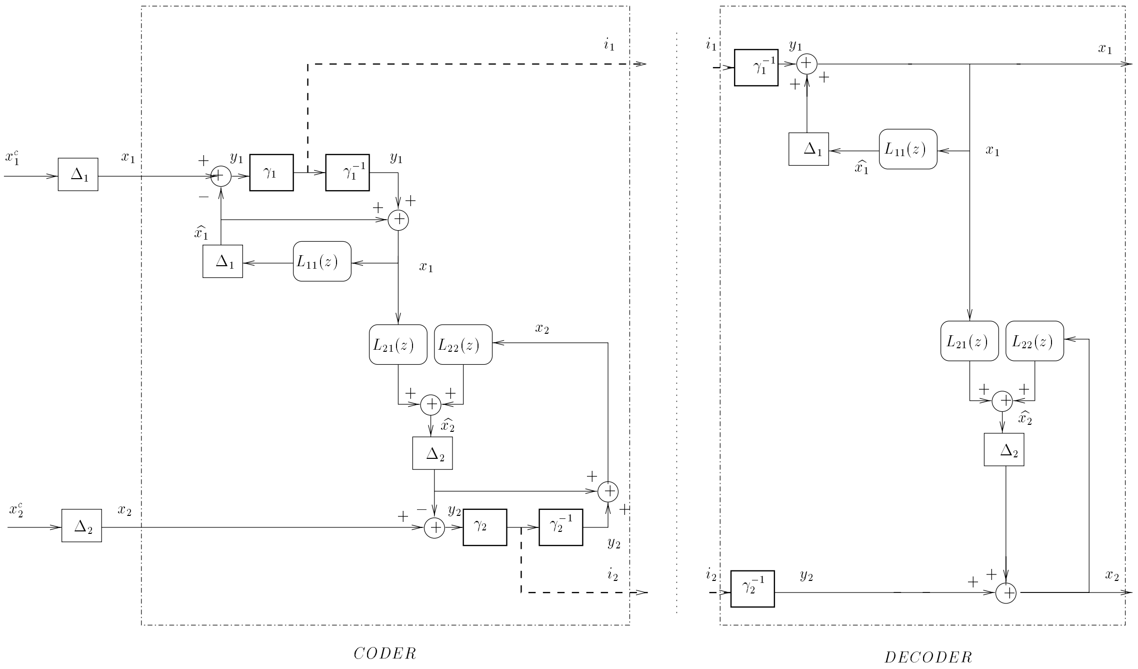


Figure 8.2: “One-shot” integer-to-integer triangular multichannel prediction for $N = 2$.

The entries of the lower triangular MIMO prediction matrix $L(z)$ (which may be written as $I - \overline{L}(z)$, where I is the identity matrix) are $L_{ij}(z)$. $L_{ij}(z)$, $i \neq j$ are Wiener filters, and $L_{ii}(z)$ are optimal causal linear prediction filters. The rounding operations denoted by Δ_i (high resolution is assumed) ensure the losslessness of the structure: each $\widehat{x}_i$ is quantized to the same multiple of Δ_i as x_i. The y_i are obtained by $y_i = x_i - [\widehat{x}_i]_{\Delta_i}$, and further (independently) entropy coded. At the decoder, the x_i are recovered by $y_i + \widehat{x}_i$.

Any lossless MIMO predictor can be written as $L_{int}^q(z) = I - \overline{L}_{int}^q(z)$. In the triangular case, only the diagonal entries of figure 8.3 are causal. In the classical MIMO prediction case, $\overline{L}_{int}^q$ is striclty causal. A generic block diagram of the “generalized” MIMO predictor is presented in figure 8.3.

8.3.2 Case of Finite Prediction Orders

An application of the classical MIMO prediction to lossless audio coding has been recently presented in [167]. In this case, FIR filters are used to remove inter- and intra-channel correlations of stereo and multichannel audio signals (16 b/s, 48kHz). The orders of these filters are adaptatively chosen (on a frame

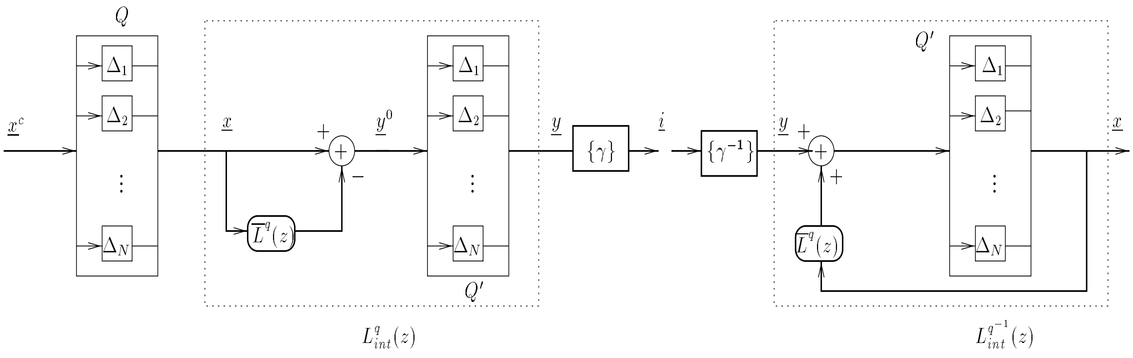


Figure 8.3: Equivalent block diagram of the “one-shot” integer-to-integer multichannel predictors.

basis of 1024 samples) among a set of possible orders (30 for the intra-signal filters and 10 for the (causal) intersignal filters). Those orders are retained which minimize the bitrate. Finding the optimal order combinations results in a great complexity, even for stereo signals. Strategies are thus proposed to reduce this complexity. The orders of the intrasignal filters $L_{ii}(z)$ are determined by using Levinson algorithms. Once these orders are fixed, the best order for the causal crossband predictor $L_{ii}(z), i \neq j$ is evaluated. Further complexity reduction can be achieved by increasing all the orders simultaneously. After the optimization procedure, the coefficients are quantized with 12 bits each, and transmitted to the decoder. The results show that appreciable bitrate reduction may be achieved by these techniques. They are also interesting in the sense that they show how the compression efficiency depends on a carefull compromise of the orders w.r.t. the complexity, and the quantization accuracy. A success of the structure relies on the decorrelation efficiency, which in turn relies on positioning judiciously the taps of the filters. As discussed in chapter 5, the triangular MIMO lossless predictor may be useful in this framework, since the intersignal filters are not restricted to be causal, and some non causality may be allowed in frame-based coding schemes.

8.3.3 Coding gain

We can define the gain $G_{L(z)}$ for the lossless implementation of a transform $L(z)$ as the difference $r_{scal}(\underline{x}) - r_{scal,L(z)}(\underline{y})$, where $r_{scal}(\underline{x})$ was defined in (8.8), and $r_{scal,L(z)}(\underline{y})$ is the actual bitrate required to scalar entropy code the decorrelated transform components y_i. This gain may be written as

$$G_{L(z)} = r_{scal}(\underline{x}) - r_{scal,L(z)}(\underline{y}) = \frac{1}{N}\sum_{i=1}^{N} H(x_i) - \frac{1}{N}\sum_{i=1}^{N} H(y_i). \tag{8.10}$$

We shall now investigate the effects of the rounding operations on the compression performance. Let us denote by $y^0_{i,k}$ the optimal prediction error obtained by applying $L(z)$ to $\underline{x}$ (that is, without the rounding operations ensuring the losslessness). Then the $y_{i,k}$ can be related to the $y^0_{i,k}$ by

$$\begin{aligned} \underline{y}_k &= \underline{x}_k - [\overline{L}(q)\underline{x}_k]_{\Delta_i} \\ &= [\underline{x}_k - \overline{L}(q)\underline{x}_k]_{\Delta_i} = [\underline{y}^0_k]_{\Delta_i}, \end{aligned} \tag{8.11}$$

where $[\underline{y}^0]_{\Delta_i}$ denotes quantization with stepsize Δ_i of the ith component of $\underline{y}^0$, and the notation (q) denotes the unit delay operator. Thus, $y_{i,k}$ may be seen as the optimal prediction error $y^0_{i,k}$ quantized with the same stepsize as $x_{i,k}$. Since $\underline{y}^0$ is a totally decorrelated process, we have from (5.7)

$$\prod_{i=1}^{N} \sigma^2_{y_i^0} = \det R_{\underline{y}^0\underline{y}^0} = e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln[\det(S_{\underline{x}\underline{x}}(f))]\, df}. \tag{8.12}$$

The bitrate $r_{scal,L(z)}(\underline{y})$ may then be written as

$$\begin{aligned} r_{scal,L(z)}(\underline{y}) &= \tfrac{1}{N}\sum_{i=1}^{N} h(y_i^0) - \frac{1}{N}\sum_{i=1}^{N}\log_2 \Delta_i \\ &\approx \tfrac{1}{2}\log_2 2\pi e \left(e^{\frac{1}{N}\int_{-1/2}^{1/2} \ln\det S_{\underline{x}\underline{x}}(f)df}\right) - \tfrac{1}{N}\sum_{i=1}^{N}\log_2 \Delta_i, \end{aligned} \tag{8.13}$$

which is the generalization of the rate of the one-shot coder expressed in (6.19).
Using (8.8), (8.10) and (8.13), we get the following expression of the gain :

$$G_{L(z)} \approx \frac{1}{2N}\log_2 \frac{\det\ \mathrm{diag}\{R_{\underline{x}^c\underline{x}^c}\}}{e^{\int_{-1/2}^{1/2} \ln\det S_{\underline{x}\underline{x}}(f)df}}. \tag{8.14}$$

In the case of equal $\Delta_i = \Delta_{VHR}$, expression (8.14) may be approximated, similarly to (6.21), by

$$G_{L(z)} \approx \underbrace{\frac{1}{2N}\log_2 \frac{\det\ \mathrm{diag}\{R_{\underline{x}^c\underline{x}^c}\}}{e^{\int_{-1/2}^{1/2} \ln\det S_{\underline{x}^c\underline{x}^c} df}}}_{G_{Max}} - \underbrace{\frac{\Delta^2_{VHR}}{24N\ln 2}\left(\int_{-1/2}^{1/2} \mathrm{tr}\{\ S^{-1}_{\underline{x}^c\underline{x}^c}(f)\}df\right)}_{Excess\ bitrate\ due\ to\ the\ lossless\ constraint}, \tag{8.15}$$

where tr stands for the trace operator.
This gain is achieved by any optimal decorrelating approach. Thus, in the case of very high resolution, vector entropy coders performance can be approached by an optimal MIMO lossless prediction followed by scalar entropy coders. Comparing with the lossless implementation of the LDU in figure 6.4 of chapter 6, note that the quantization stage Q' involves N quantizers instead of $N-1$, because of the presence in MIMO prediction of the intrasignal prediction filter $L_{11}(z)$. This renders the excess bitrate caused by the lossless constraint the same for all decorrelation approaches.

8.4 Two-Stage MIMO prediction

8.4.1 Structure

We will now investigate the compression performance of multiresolution approaches based on the decorrelating transform $L(z)$. For these approaches, a uniform quantizer Q_1 is introduced in the (A)DPCM prediction loops, whose effect is to reduce the entropy of the transform signals y_i^q. These signals represent low resolution versions of the transform signals y_i described in the previous section. The error signals e_i, $i = 1, ..., N$, are then generated by substraction, and separately entropy coded. Note that the transform signals are computed by substracting the optimal estimate of x_i based on the past *quantized* samples x_i^q, and by quantizing with stepsize Δ_i the resulting error prediction[2]. Thus, only the available x_i^q at the decoder should be used to compute the remaining $x_j^q,\ j > i$. A two-stage structure based on the triangular MIMO predictor is depicted in figure 8.4, for $N = 2$.

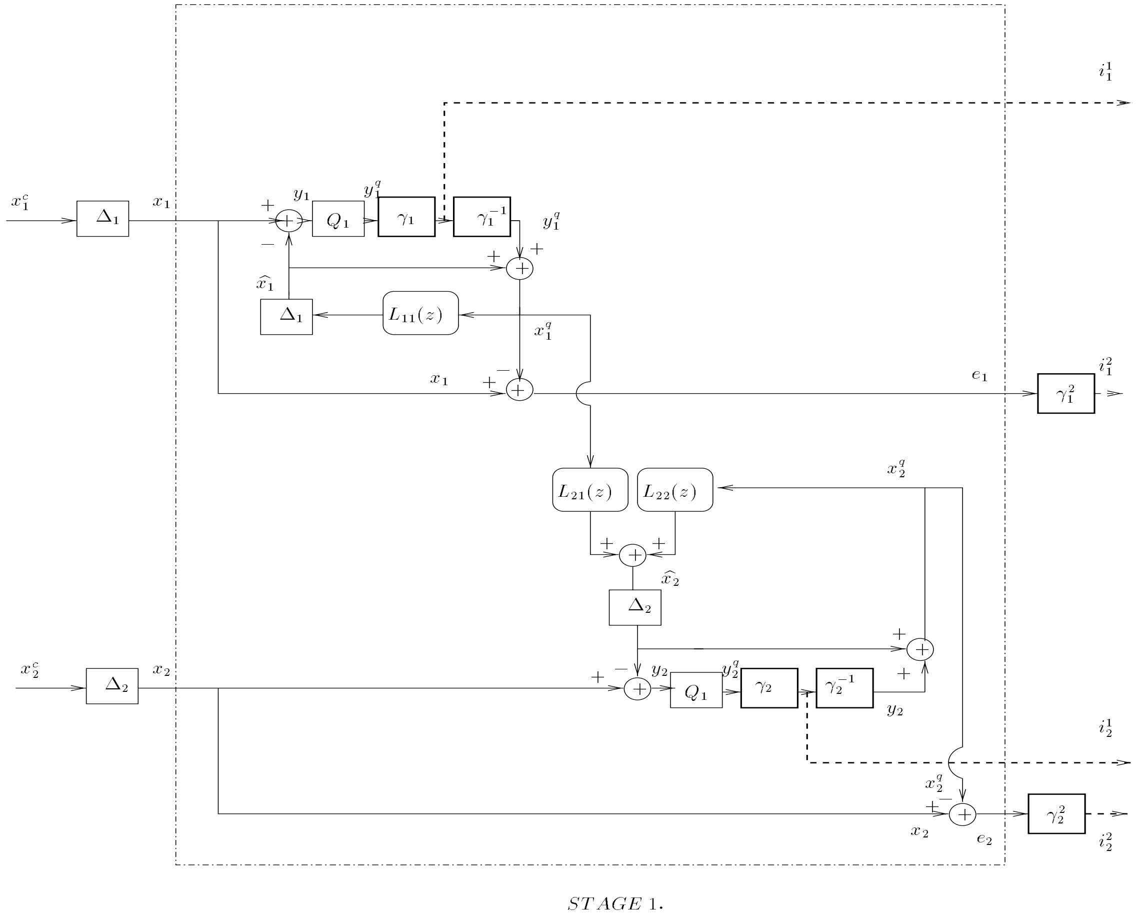


Figure 8.4: Two-stage encoder of the scalable lossless multichannel triangular predictor, for $N = 2$. The bitrates for $\{i_{1:N}^1\}$ and $\{i_{1:N}^2\}$ are fixed by the quantizer Q_1.

This structure resembles the embedded DPCM coders of [124, 127], evocated in the introduction to the second part of this thesis. In these schemes, the predictions are also based on *core* bits (lossy versions x_i^q), and may be seen as lossless, multichannel, and possibly noncausal version of these algorithms. The overall bitrate is the average r_{LR} of the bitrates corresponding to the low resolution substreams $\{i_k^1\}$, $k = 1, ..., N$, plus the average $\bar{r}$ of the rates corresponding to $\{i_k^2\}$, $k = 1, ..., N$ (substreams of the error signals). In order to simplify the derivations, we assume in this section that the Δ_i corresponding to the preliminary quantization stage are all equal, $\Delta_i = \Delta_{VHR}$. Moreover, we assume w.l.g. that the variances $\sigma_{x_i}^2$ are large

[2] The prediction is computed by means of quantized data because we are interested in a low resolution signal which can be computed *independently* of the error signals.

in comparison with 1, and that $\Delta_{VHR} = 1$. Thus, x_i are integer valued, and $H(x_i) \approx h(x_i^c) - \log_2 \Delta_i \approx h(x_i^c)$, and $S_{\underline{x}\underline{x}}(f) \approx S_{\underline{x}^c\underline{x}^c}(f)$. This is equivalent to neglecting the effects of Δ_{VHR} w.r.t. those of Δ_{Q_1}. The stepsize Δ_{Q_1} is generally much larger than 1: for example, if one wishes to divide by 2 the 0th order entropy of an integer-valued source with variance 10^4, the corresponding Δ_{Q_1} is $\approx 20 \gg \Delta_{VHR} = 1$.

8.4.2 Analysis of the Rates

We shall now analyze the bitrate dedicated to the low resolution version $r_{LR} = \frac{1}{N}\sum\limits_{i=1}^{N} H(y_i^q)$. Considering the figure 8.4, each $y_{i,k}$ is the optimal prediction of $x_{i,k}$ based on the past quantized value of x_i, and on all the quantized components of x_j, for all $j < i$. (For the classical MIMO predictor, the corresponding $y_{i,k}$ are based on the past and current quantized samples $\underline{x}^q_{1:i-1,-\infty:k}$, and on all the past samples $\underline{x}^q_{i,-\infty:k-1}$). Assuming that the y_i are Gaussian, we have

$$r_{LR} = \frac{1}{N}\sum_{i=1}^{N} H(y_i^q) \approx \frac{1}{2N}\log_2(2\pi e)^N \prod_{i=1}^{N}\sigma^2_{y_i} - \frac{1}{N}\sum_{i=1}^{N}\log_2 \Delta_{Q_1}. \tag{8.16}$$

We now use the result (5.23) from chapter 5. We apply $L(z)$ to decorrelate the vectorial source $\underline{x}$ in closed loop around quantizers with stepsize Δ, that is, by computing the predictions by means of quantized data of $\underline{x}$. The resulting vectorial process is $\underline{y}$. Then the variances of the process $\underline{y}$ can be approximately related to the variances $\sigma^2_{y_i^0}$ of $\underline{y}^0$ (eq. (8.12), fig. 8.3), and to $S_{\underline{x}^c\underline{x}^c}(f)$ by

$$\prod_{i=1}^{N}\sigma^2_{y_i} \approx \prod_{i=1}^{N}\sigma^2_{y_i^0}\left(1 + \frac{\Delta_Q^2}{12}\left[\int_{-\frac{1}{2}}^{\frac{1}{2}} \mathrm{tr}\, S^{-1}_{\underline{x}^c\underline{x}^c}(f)df - \sum_{i=1}^{N}\frac{1}{\sigma^2_{y_i^0}}\right]\right). \tag{8.17}$$

Applying (8.17) to (8.16) yields

$$\begin{aligned} r_{LR} &\approx \tfrac{1}{2}\log_2 2\pi e^{\frac{1}{N}\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln\det S_{\underline{x}^c\underline{x}^c}}\left(1 + \frac{\Delta^2_{Q_1}}{24N\ln 2}\left[\int_{-1/2}^{1/2} \mathrm{tr}\, S^{-1}_{\underline{x}^c\underline{x}^c}(f)df - \sum_{i=1}^{N}\frac{1}{\sigma^2_{y_i^0}}\right]\right) - \log_2\Delta_{Q_1} \\ &\approx r_{scal,L(z)}(\underline{y})\left(1 + \underbrace{\frac{\Delta^2_{Q_1}}{24N\ln 2}\left[\int_{-1/2}^{1/2} \mathrm{tr}\ S^{-1}_{\underline{x}^c\underline{x}^c}(f)df - \sum_{i=1}^{N}\frac{1}{\sigma^2_{y_i^0}}\right]}_{Factor\ Excess\ bitrate\ due\ to\ noise\ feedback}\right) \underbrace{-\log_2\Delta_{Q_1}}_{Bitrate\ reduction\ due\ to\ Q_1}, \end{aligned} \tag{8.18}$$

Minimizing this excess bitrate entails maximizing $\sum\limits_{i=1}^{N}\frac{1}{\sigma^2_{y_i^0}}$. From the theorem of chapter 5, this in turn entails processing the signals in order of decreasing variance. Moreover, this excess bitrate will be minimized by the lossless triangular MIMO predictor.

Now, the bitrate $\overline{r}$ dedicated to the error signals, corresponds to the entropies of the r.v.s $e_i = x_i - x_i^q$, which were calculated in chapter 7, eq. (7.11). Thus, depending on the parity of Δ_{Q_1}, we obtain

$$\begin{aligned} \overline{r}_{even} &\approx \log_2\Delta_{Q_1} + \tfrac{1}{\Delta_{Q_1}}, \\ \overline{r}_{odd} &\approx \log_2\Delta_{Q_1}. \end{aligned} \tag{8.19}$$

8.5 Mutistage Integer-to-Integer Multichannel Prediction

Finally, one may elaborate multiresolution structures based on M two-stage lossless coders. These schemes allow one to split the rate obtained by a one-shot coder $r_{scal,L(z)}(\underline{y})$ into $M+1$ substreams with rates $r_i,\ i=1,\cdots,M+1$. These rates are controled by the stepsizes of Δ_{Q_i} of each two-stage block, see figure 8.5. For the stages $i>2$, the predictors become useless if the error signals are white.

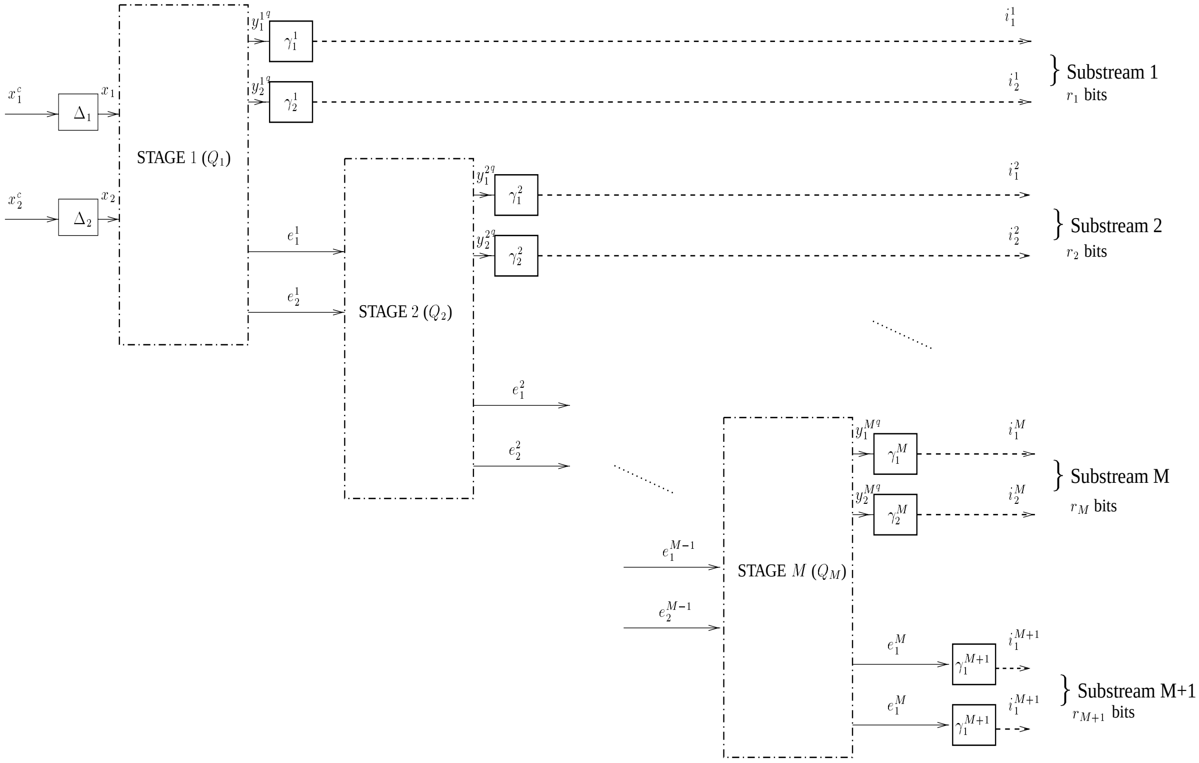


Figure 8.5: Multistage structure of the lossless multichannel (triangular) prediction scalable encoder for $N=2$. The bitrates of the substreams are determined by the quantizers Q_i.

Suppose we dispose of partially- or un-compressed data $\underline{x}$. Suppose we wish to transmit the data by means of $M+1$ substreams corresponding to different resolution levels with imposed rates R_i ($\sum_{i=1}^{M+1} R_i \approx r_{scal,L(z)}(\underline{y})$). How should we choose the stepsizes Δ_{Q_k} of the M uniform quantizers ?

For the sake of simplicity, we will neglect the term corresponding to the noise feedback in (8.18), and assume odd, and sufficiently large stepsizes Δ_{Q_i}.

In a first step, the minimum bitrate $r_{scal,L(z)}(\underline{y})$ (8.13) is obtained by compressing the data with some one-shot lossless coder. Now, the two-stage structure of figure 8.4 will yield a first substream with rate

$$r_1 = r_{LR} \approx r_{scal,L(z)}(\underline{y}) - \log_2 \Delta_{Q_1}, \tag{8.20}$$

and a complementary susbstream with rate $\overline{r} \approx \log_2 \Delta_{Q_1}$. If we use a second stage, the previous error signal with rate $\overline{r}$ will be divided into two substreams with rates $r_2 \approx \log_2 \frac{\Delta_{Q_1}}{\Delta_{Q_2}}$, and $\overline{r}_2 \approx \log_2 \Delta_{Q_2}$. Thus, a structure using M stages will yield a first substream with rate r_1 given by (8.20), $M-1$ complementary substreams with rates

$$r_j \approx \log_2 \frac{\Delta_{Q_{j-1}}}{\Delta_{Q_j}},\ j = 2, 3...M, \tag{8.21}$$

and a last substream with rate $r_{M+1} \approx \log_2 \Delta_{Q_M}$.

It can easily be checked that the constraint $r_1 \approx R_1$ imposes $\Delta_{Q_1} \approx [2^{r_{scal,L(z)}(\underline{y})-R_1}]_1$. Similarly, the constraints $r_k \approx R_k$ impose $\Delta_{Q_k} \approx [\Delta_{Q_{k-1}} 2^{-R_k}]_1$, for $k = 2, ...M$. Thus, the stepsizes Δ_k of the M uniform quantizers should be determined by the simple rule of thumb

$$\Delta_{Q_k} \approx \left[2^{r_{scal,L(z)}(\underline{y}) - \sum_{i=1}^{k} R_i} \right]_1,\quad k = 1, ..., M. \tag{8.22}$$

8.6 Numerical Results

Some numerical results regarding the strategy (8.22) are presented in this section. We implemented the structure of figure 8.5 for the multiresolution coding of a two dimensional memoryless vector source. In this case, the temporal decorrelation becomes useless, and each two-stage block reduces to the structure presented in figure 7.3, where the Δ_i are equal, and determined, for each block, by the rule (8.22). The covariance matrix of the source was $R = HR_{AR_1}H^T$ with diagonal elements 1.6×10^4, and 10^4. Each vector was quantized with stepsize $\Delta_{VHR} = 1$. The resulting theoretical bitrate for the corresponding single-stage coder $r_{scal,L(z)}(\underline{y})$ is given by 8.13. We chose to compress this source by means of three substreams with rates R_1, and $R_2 = R_3$. For different target combinations, the stepsizes Δ_{Q_1} and Δ_{Q_2} were fixed according to (8.22) (with the restriction to be odd). The resulting rates were measured either by the entropy (fig. 8.6), or by the average rate obtained by Huffman codes (fig. 8.7), for sequences of length 5×10^4. The rate $r_{scal}(\underline{x})$, obtained without compression, is plotted in full line. The correspondence of the stepsizes for each combination of target rates is given in the table below.

Combination	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Δ_{Q_1}	1	1	1	1	3	3	3	5	5	7	9	9	13	15	19	23	27	35	43
Δ_{Q_2}	1	1	1	1	3	3	3	3	3	3	3	3	3	5	5	5	5	7	7

It can be observed that the bitrates actually delivered by the multiresolution structure match approximately the target ones when the stepsizes become large w.r.t. $\Delta_{VHR} = 1$ (cases where the rate of the low resolution version is decreased by more than 20% w.r.t. the rate of the single-stage structure).

8.7 Conclusions

This chapter dealt with optimal lossless coding of vectorial signals. The coding structures investigated in a first step involved single-stage structures using prediction matrices $L(z)$ of the generalized MIMO prediction framework. The corresponding compression performance were compared to the optimal compression performance. The particular cases of the classical and the triangular MIMO predictors were investigated, and shown to present equivalent performance. In a second step, we investigated the performance of two-stage structures where ADPCM loops were introduced. The quantizers of these loops allow one to choose the respective bitrates for both the error and the low resolution signals. For these two-stages structures, the overall bitrate delivered by the multiresolution structure was compared to that of the corresponding "one-shot" approach. These two-stages structures were shown to be slightly suboptimal because of the noise feedback created in ADPCM loops. Finally, we showed that the two-stage structure could easily be extended to M stages. A strategy was proposed so that the delivered bitrates approach some predetermined target rates. This strategy is efficient if the rate of the low resolution signal is sufficiently decreased w.r.t. the overall rate.

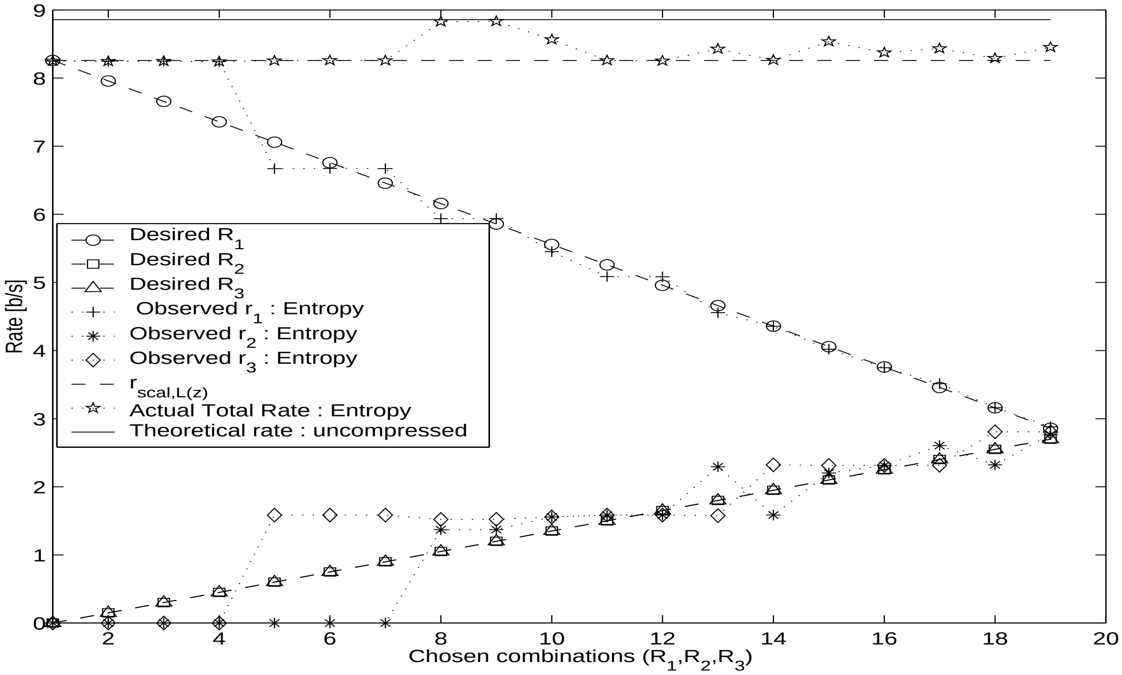


Figure 8.6: Actual entropies delivered by the multistage structure *vs* several combinations of target rates.

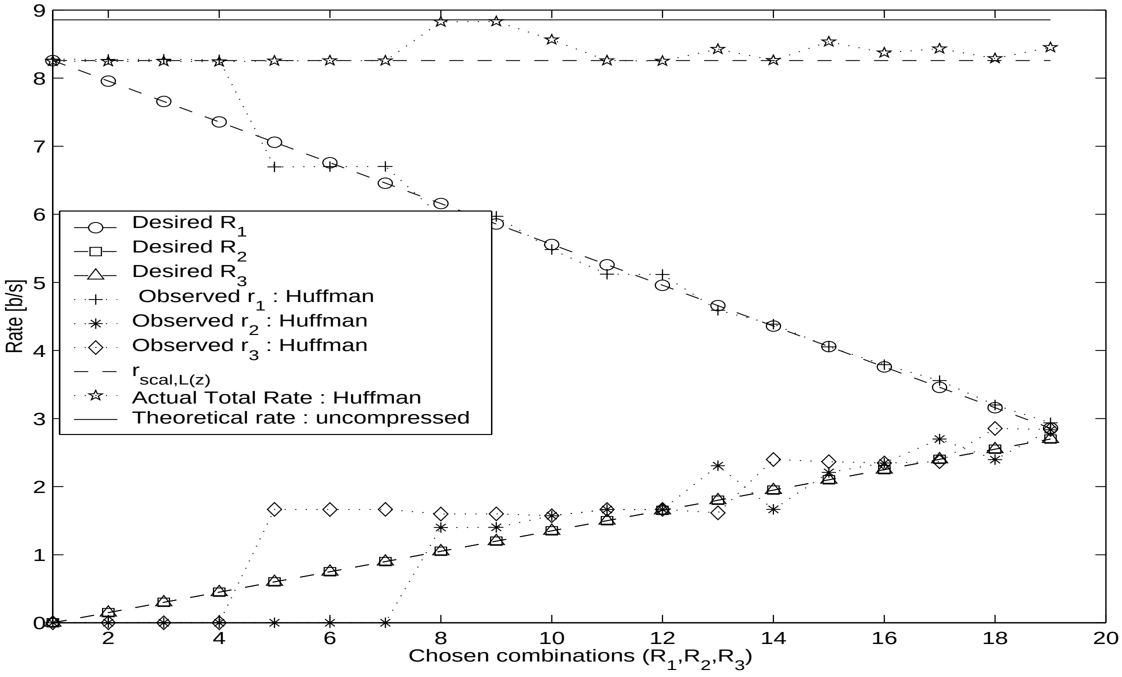


Figure 8.7: Actual rates obtained by Huffman coding from the multistage structure *vs* several combinations of target rates.

Chapter 9

Conclusions

This thesis has presented various coding structures derived from a general causal framework. As far as the origin of these results is concerned, one may recall that this framework is issued from an analysis-by-synthesis structure based on a Laplacian Pyramid.

The performances of the corresponding coding systems were analyzed in both the lossy and the lossless coding frameworks. In the following, we summarize the main results of the thesis, and outline then open problems and further works.

In a transform coding framework firstly, we showed that the proposed causal transform performs a Lower-Diagonal-Upper factorization of the covariance matrix of the vectorial source to be coded. It is not unitary but causal, and is based on optimal prediction. A theoretical analysis showed that in the limit of high rates, this transform achieves the same performance as the KLT, which is the optimal transform for Gaussian sources. As a consequence of its non-orthogonality, we showed that efficient causal coding structures should be implemented in closed loop around the quantizers, as in DPCM systems. We proposed a general analysis of the corresponding noise feedback for both systems working at high rates, and for particular systems using entropy coded uniform quantizers with equal quantization stepsizes. For these systems, we showed that the causal transform competes with the KLT at average bitrate budgets higher than 2.5 b/s. As the KLT, the LDU is data dependent, and should thus be updated in case of changes in the source statistics. This led us secondly to turn our investigations to backward adaptive transform coding systems. The first attempt to model theoretically the performances of the causal and unitary transforms in this context consisted in analyzing the corresponding perturbation effects w.r.t. to the classical transform coding framework. In order to make tractable analyses, several simplifying assumptions were made. The proposed model match accurately the corresponding idealized coding systems. We turned then to three practical backward adaptive transform coding schemes, including fixed and adaptive stepsizes. The proposed analyses suggest that the corresponding algorithms are universal in the sense that the transforms converge to the optimal transforms for sources among a given class. In the case where both the stepsizes and the transforms are adaptive, the algorithm using a Sheppard's correction on the second order moment estimates converge to the target stepsize, and thus, to the target distortion at high rates. The proposed models match the actual convergence process for rates higher than approximately 2.5 b/s.

We then considered optimal coding of vectorial signals. We showed in this case that the optimal causal decorrelating scheme could still be described by a triangular prediction matrix whose entries are optimal prediction filters. The diagonal filters are scalar intrasignal prediction filters, and the off-diagonal predictors are Wiener filters performing the intersignal decorrelation. This decorrelating scheme led to the notion of "generalized" MIMO prediction, in which a certain degree of non causality may be allowed for the off-diagonal prediction filters. Previously introduced MIMO decorrelation approaches were shown to be special cases of this description, namely the classical, and the triangular MIMO predictors. For the latter, the " causality" between channels becomes processing the channels in a certain order; some signals may be coded using the coded/decoded versions of the "previous" signals. We then showed that if the quantization noise feedback is taken into account, the optimal strategy is to decorrelate the signals by order of decreasing variance. Moreover, the triangular predictor was shown to be the most efficient predictor.

The second part of this thesis analyzed the performances of causal approaches in a lossless coding framework. Our results regard integer-to-integer transforms, and multiresolution structures.

For single-stage structures, the bitrate reduction operated by a lossless coding scheme was defined as a lossless coding gain. An upper bound for this gain was expressed in terms of mutual information shared by the random variables to be coded. The inherent suboptimality of integer-to-integer transforms was then compared for the LDU and the KLT. Finally, adaptive single-stage lossless transform coding systems were investigated. For a fixed number of vectors K, we evaluated, for both the causal and the unitary cases , the bitrate reduction that could be achieved by the corresponding estimated transform. We showed that for single-stage systems, the respective performances of the LDU *vs* those of the KLT are reversed w.r.t. the classical transform coding case. The integer-to-integer KLT achieves the same compression as that of the lossless LDU in the limit of high rates only. At lower rates, the KLT's compression performances are more deteriorated by the integer-to-integer constraint than those of the LDU, because the KLT is not triangular.

We then studied two-stage structures based on the KLT and on the LDU transform. For a fixed preliminary quantization stage (and for a sufficiently high resolution), we analyzed the bitrate required to entropy code the corresponding low resolution and error signals. The resulting overall bitrate was compared to that obtained with the corresponding single-stage structure. We showed that while orthogonal transforms tend to "gaussianize" the error signals, the LDU benefits from keeping them uniform. As a consequence, the orthogonal transforms, including the KLT, were shown to be approximately 0.25 b/s/ch suboptimal w.r.t. their causal counterpart. Finally, we underlined several other practical coding advantages of the LDU, namely the ability of switching easily from a single- to a multi-stage structure, and that of allowing one to represent the different channels with different resolution levels. Moreover, we showed that the errors in the causal case can be made equally likely, which makes the entropy coding very straightforward.

Finally, we applied our results about optimal coding of vectorial signals to the frameworks of the single- and multi-stage lossless structures described so far. The coding structures investigated in a first step involved single-stage structures using prediction matrices $L(z)$ of the generalized MIMO prediction framework. The corresponding compression performances were compared to the optimal compression performances, as achievable by any lossless coding technique. The particular cases of the classical and the triangular MIMO predictors were investigated, and shown to present equivalent performances. In a second step, we investigated the performances of two-stage structures where (A)DPCM loops were introduced. The quantizers of these loops allow one to choose the respective bitrates for both the error and the low resolution signals. For these two-stages structures, the overall bitrate delivered by the multiresolution structure was compared to that of the corresponding “one-shot” approach. These two-stage structures were shown to be slightly suboptimal because of the noise feedback created in the (A)DPCM loops. Finally, we showed that the two-stage structure could easily be extended to a larger number of stages. In that case, a simple method was proposed so that the delivered bitrates approach some predetermined target rates. This method is efficient if the rate of the low resolution signal is sufficiently decreased w.r.t. the overall rate.

As can be seen from the summary of these results, various coding techniques appeared in the scope of the proposed investigations[1], including transform coding, subband coding, integer-to-integer transforms, multiresolution coding, and combinations thereof. The choice of such a wide scope is of double value. On the one hand, this choice was necessary to describe the versatile forms of the causal coding approach. On the other hand, this choice led us to let for further work some interesting topics, which were only evoked, or taken up in passing throughout the developments . Some of the presented analyses were focused on a statistical modeling of the coding performances of particular causal systems; these systems may, however, be further elaborated for the purposes of particular applications, *e.g.* audio coding. In particular, it seems interesting to investigate the performances of the backward adaptive LDU or the triangular MIMO predictor for multichannel audio sources, with appropriate and possibly time-varying adaptation windows. As for the triangular MIMO predictor, the coding efficiency will also rely on a careful positioning of the taps of the crossband filters. The degrees of noncausality allowed to these filters should be optimized w.r.t. the framelength, or w.r.t. some reconstruction delay between the different channels for a sample-by-sample coding scheme. Besides, perceptual considerations, which were not mentionned throughout the thesis, may be accounted for by introducing noise shaping filters, as in classical scalar (A)DPCM. This technique would regard both the lossy encoder of the triangular predictor, and the low resolution signal of the corresponding

[1] In addition to source coding, the generalized MIMO prediction have interesting applications in multiuser detection [53].

lossless multi-stage encoder.

The proposed bandwidth expansion operated by the Wiener filters may also be improved by optimizing the analysis filters (on which depend the information shared by the subbands), and by carefully optimizing the number of coefficients of the filters transmitted to the decoder.

Finally, one may attempt to extend the presented theoretical results established in the Gaussian case to different sources. One may consider Gaussian mixture models, which allow to model sources with arbitrary probability density functions.

Chapter 10

Résumé Détaillé en Français

10.1 Introduction

La nécéssité de “comprimer” les signaux numériques trouve son origine dans les moyens limités dont disposent les communications numériques : la compression permet d’économiser la bande passante des canaux sans-fil ou internet; elle permet aussi d’économiser l’espace mémoire en ce qui concerne leur stockage. D’une façon générale, le codage de source consiste à mettre au point des techniques permettant, suivant l’application visée, de déterminer le meilleur compromis entre la qualité avec laquelle les informations seront représentées, et la ressource, ou le débit, qui sera nécessaire pour décrire la représentation choisie. Selon que l’information initiale peut être partiellement, ou parfaitement reproduite après l’opération de codage, on parle de codage avec, ou sans perte. Cette thèse présente diverses techniques, et l’évaluation de leur efficacité, pour ces deux types de codage.

L’ *information* considérée dans cette thèse sera représentée par des signaux vectoriels, qui forment une large classe de signaux, incluant par exemple les signaux scalaires ou les signaux multicanaux. Ces derniers peuvent être construits dès que plusieurs signaux scalaires sont, pour des applications diverses, regroupés. Dès lors que les signaux scalaires individuels présentent des dépendances, comme certains signaux audio par exemple, il y a un intérêt à les traiter conjointement, en vue d’une compression plus efficace.

L’idée initiale de développer des techniques adaptées aux signaux audio[1] a motivé ce choix d’une représentation vectorielle. Bien que quelques applications soient présentées pour ce type de signaux, l’hypothèse de signaux gaussiens est souvent retenue. Les sources Gaussiennes ont un statut particulier en théorie de l’information. Shannon [25] a montré qu’une source Gaussienne indépendante et identiquement distribuée (i.i.d.) possède la fonction débit-distorsion la plus défavorable, comparativement à n’importe quelle source i.i.d. de même variance, montrant par là que les Gaussiennes constituent un extremum du point de vue du codage de source. Historiquement, ce constat a fourni les éléments pour élaborer des techniques de *quantification robuste*[51]. Par ailleurs, pour une source de densité de probabilité arbitraire, on peut utiliser avantageusement le théorème de la limite centrale et un code construit pour une Gaussienne [52]. Toutefois, on ne prétend pas utiliser ici le modèle de source Gaussienne pour fournir des approches de quantification robuste ou des méthodes de codage de sources arbitraires par préfiltrage et quantificateurs Gaussiens. Cette hypothèse permet surtout d’obtenir des résultats analytiques relatifs schémas de codage considérés, de les comparer et de prouver, le cas échéant, leur optimalité. Dans ce sens, elle fournit un cadre de travail adapté aux investigations théoriques préliminaires associées aux schémas de codage présentés.

Nous inspirant de [19] et [20], cette thèse aurait aussi pu être intitulée “Variations on a causal coding theme”: le thème de la *causalité* dans le codage de source est le lien essentiel entre les chapitres de cette thèse[2]. Plusieurs schémas de codage causaux sont présentés et analysés au long du document. Dans tous les cas où le schéma de codage comprend une transformation matricielle (à coefficients scalaires) causale,

[1] Les premiers résultats de ce travail ont été obtenus dans le cadre du projet RNRT *COBASCA* : COdage en Bande élargie avec partage Adaptatif du débit entre Source et CAnal pour Réseaux cellulaires de deuxième et troisième générations (UMTS).

[2] Nous avons néanmoins essayé de faire en sorte que les chapitres puissent être lus indépendamment, et avons repris, quand cela semblait nécéssaire, les résultats précédemment établis.

nous en comparons les performances avec le schéma équivalent basé sur une transformation optimale pour les sources gaussiennes, la transformation de Karhunen-Loève [42, 43] (Karhunen-Loève Transform, KLT).

Cette thèse comprend deux parties. La première traite du codage avec pertes (ou compression), et la deuxième du codage sans pertes (ou compaction). Chaque partie comporte une introduction détaillée présentant la problématique et la trame des divers développements. Un résumé est présenté au début de chaque chapitre.

Après un chapitre d'introduction, rappelant les principaux concepts et définitions de théorie de l'information nécéssaires au codage de source, la première partie de cette thèse concerne le codage par transformée (CT). Le CT peut apparaître, d'un point de vue théorique comme pratique, comme une technique parfaitement maîtrisée et aboutie. Un des buts de cette partie est de montrer que des innovations majeures sont encore possibles dans ce domaine. Dans le cadre du CT standard tout d'abord, ces innovations concernent l'introduction d'une transformation qui n'est pas unitaire mais causale, et qui présente des performances comparables à celle de la KLT. Par la suite, les apports théoriques de cette première partie concernent un domaine presque totalement inexploré, celui du codage par transformée en boucle fermée, ou “en ligne”, ou encore sans ”side-information”.

Dans la fin de cette première partie, la transformation matricielle causale est généralisée au cas où les coefficients de la matrice de transformation triangulaire sont des filtres prédicteurs (prédiction MIMO, Multi Input Multi Output, triangulaire). Cette généralisation débouche sur la prédiction MIMO dite “généralisée”, pour indiquer que la prédiction MIMO classique et la prédiction MIMO triangulaire constituent deux cas particuliers, parmi une infinité, d'une même approche totalement décorrélatrice, et “causale” dans un sens plus large. Un bref historique des principaux résultats est présenté en fin de partie.

La seconde partie de cette thèse présente et analyse des techniques de codage causales et sans pertes, dérivées des structures présentées dans la première partie.

Les thèmes de cette partie sont premièrement les transformations d'entiers à entiers, qui peuvent être vue comme une analogie “sans pertes” (et non linéaires) du codage par transformée, et qui ont été récemment l'objet de nombreux travaux. Dans ce cadre, la transformation causale présente aussi une alternative intéressante aux transformations habituellement utilisées (unitaires). Le deuxième thème récurrent dans cette seconde partie est le codage multirésolution qui permet, en augmentant le débit apporté à un premier codage grossier d'une source, d'en améliorer la représentation. Par ailleurs, le codage sans pertes de signaux audio multicanaux est actuellement un terrain de recherches actives, et les résultats proposés s'appliquent naturellement à ce domaine. Enfin, les résultats et les structures présentées peuvent être appliqués au codage de l'image également.

La structure de cette seconde partie ressemble à celle de la première: les deux premiers chapitres couvrent des techniques liées à des transformations matricielles à coefficients scalaires; la dernière partie généralise ces derniers résultats dans le cas où la transformation décorrélante est sans perte, et basée sur un filtrage matriciel causal de type MIMO généralisé.

10.2 Première Partie: Codage Causal avec Pertes

10.2.1 Introduction

Le codage par transformée est populaire parce qu'il permet un compromis attratif entre la complexité et les performances. Cette technique est largement analysée et commentée dans la littérature, et les systèmes de codage de source qui utilisent ce genre de code sont innombrables. Il existe de nombreuses transformations, qui présentent des compromis différents entre l'efficacité théorique et des critères d'utilisation pratiques. Par efficacité théorique, on entend la capacité de décorrélation, et de compaction; des critères pratiques sont la complexité de calcul et d'implémentation de la transformation, ou des critères subjectifs liés au comportement de la transformation par rapport à la nature des signaux auxquels elle est appliquée. Le monopole du codage par transformée est détenu par les transformations orthogonales, parce qu'elles garantissent que le bruit de quantification n'est pas amplifié quand on passe du domaine transformé (vecteurs $\underline{y}$) au domaine signal (vecteurs $\underline{x}$). Parmi ces transformations, la transformation de Karhunen-Loève (KLT, Karhunen-Loève Transform) [1, 54] est traditionnellement utilisées comme parangon, parce qu'elle est optimale pour des sources Gaussiennes, quel que soit le type de quantificateurs scalaires utilisés. Un des thèmes récurrents de cette première partie est de montrer que, relativement à différents critères, les performances de la KLT peuvent être égalées par (au moins) une autre transformation, la transformation triangulaire causale dite **LDU** (Lower-Diagonal-Upper, réalisant une factorisation triangulaire de la matrice de covariance $R_{\underline{x}\underline{x}}$ du signal source).

10.2.2 Codage par Transformation Causale de Type LDU

Dans le second chapitre de cette thèse, nous dérivons la transformation causale optimale[3], dans le cadre du codage par transformée classique (hypothèses d'allocation optimale de bits, et de performances débit/distorsion constantes par rapport au débit pour les quantificateurs). De façon similaire au codage MICD (DPCM), cette transformation donne lieu à deux structures, dites en “boucle ouverte”, ou en “boucle fermée”. Comme pour le codage MICD, cette dernière est plus réaliste d'un point de vue pratique; un des schémas équivalent est représenté figure 10.1.

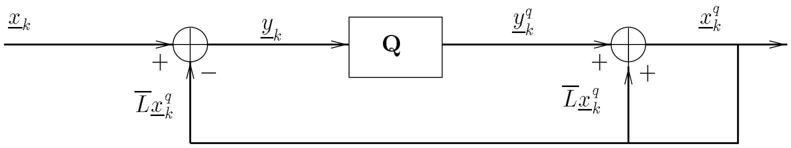


Figure 10.1: Codage par transformation causale en boucle fermée ($\mathbf{Q}$ dénote un ensemble quantificateurs scalaires).

[3] Le choix de cette contrainte de *causalité*, imposée sur la transformation, découle d'une approche de type analyse par synthèse adoptée comme axe de recherche initial, voir (10.2.6).

Notons que les erreurs de reconstruction $\underline{\widetilde{x}}_k$ et de quantification $\underline{\widetilde{y}}_k$ sont les mêmes, puisque

$$\underline{\widetilde{x}}_k = \underline{x}_k - \underline{x}_k^q = \underline{x}_k - (\underline{y}_k^q + \overline{L}\underline{x}_k^q) = \underline{x}_k - \overline{L}\underline{x}_k^q - \underline{y}_k^q = \underline{y}_k - \underline{y}_k^q = \underline{\widetilde{y}}_k. \qquad (10.1)$$

Ainsi, la conservation de l'erreur de quantification n'est pas seulement vraie en norme euclidienne, comme dans le cas unitaire, mais pour le vecteur d'erreur lui-même.

Dans un premier temps, on optimise cette transformation afin de minimiser l'erreur de reconstruction en négligeant le fait que le vecteur de référence $\overline{L}\underline{x}_k^q$ soit construit à partir de données quantifiées (hypothèse de résolution infinie). On obtient une transformation de la forme

$$L = \begin{bmatrix} 1 & & & \\ \star & \ddots & \mathbf{0} & \\ \vdots & \ddots & \ddots & \\ \star & \cdots & \star & 1 \end{bmatrix},$$

où les $\star$ représentent les coefficients de prédiction optimaux. En d'autres termes, L est telle que

$$LR_{\underline{xx}}L^T = R_{\underline{yy}} = \overline{\text{diag}}\,\{\sigma_{y_1}^2 \cdots \sigma_{y_N}^2\}, \qquad (10.2)$$

où $\overline{\text{diag}}\,\{\underline{a}\}$ représente une matrice diagonale, de diagonale $\underline{a}$. Comme chaque erreur de prédiction y_i est orthogonale aux sous-espaces générés par les $\underline{x}_{1:i-1}$, les coefficients transformés y_i sont orthogonaux, et $R_{\underline{yy}}$ est diagonale. Il suit

$$R_{\underline{xx}} = L^{-1}R_{\underline{yy}}L^{-T}, \qquad (10.3)$$

qui représente la décomposition LDU de $R_{\underline{xx}}$. On montre que puisque la matrice $R_{\underline{xx}}$ est définie positive, cette transformation existe toujours.

On montre ensuite que le gain de codage correspondant $G_L^{(0)}$, qui représente le facteur par lequel la distorsion est diminuée grâce à la transformation, est le même que celui de la KLT (dans les deux cas, la distorsion est proportionnelle à $\det R_{\underline{yy}}$; la KLT et la LDU étant unimodulaires, $\det R_{\underline{yy}} = \det R_{\underline{xx}}$ dans les deux cas).

Dans un deuxième temps, nous proposons des analyses des effets du bruit de quantification sur le gain de codage. Réoptimisant la transformation sous les hypothèses classiques du CT d'abord, et menant une analyse des perturbations au premier ordre, nous obtenons comme expression pour le gain de codage

$$G_L^{(1)} = \frac{\mathrm{E}\,||\widetilde{x}_k||^2}{\mathrm{E}\,||\widetilde{y}_k||_L^2} \approx G_L^{(0)}\left(1 - \frac{1}{N}\sigma_q^2\sum_{i=1}^{N}\frac{||\overline{L}_i||^2}{\sigma_{y_i}^2}\right), \qquad (10.4)$$

où la notation $^{(1)}$ dénote la présence du bruit de quantification sur le vecteur de référence, N est la dimension du vecteur, σ_q^2 est la variance du bruit de quantification dans le cas idéal, et λ_i sont les valeurs propres de $R_{\underline{xx}}$.

Finalement, nous analysons un système de CT pratique qui utilise des quantificateurs scalaires uniformes suivis d'un codage entropique (Entropy Coded Uniform Quantizers, ECUQ). Ces quantificateurs présentent l'avantage de réaliser simplement une allocation de bits proche de l'optimalité, en choisissant des pas de quantification égaux. Dans ce cas, les résultats théoriques et numériques montrent que la KLT et la LDU présentent des performances égales pour des débits aussi bas que 2.5 b/s. Ces résultats ont été présentés à [55, 56].

Comme la KLT, la LDU dépend des données et devrait donc être continûment adaptée aux changements de statistique de la source. Afin d'éviter le surcroît de débit associé à la transmission des paramètres de codage au décodeur, on peut chercher à adapter ces schémas sur la base des données précédemment quantifiées. Ceci pose le problème d'adaptation "en ligne" pour le CT. La faisabilité et l'évaluation des performances du CT "en ligne" est l'objet des deux chapitres suivants.

10.2.3 Analyse Haute Résolution de Schémas Idéalisés de Codage par Transformée "en ligne"

Une première contribution à la modélisation théorique de schémas CT adaptés "en ligne" consiste à analyser la perturbation par rapport au cas idéal où la matrice de covariance est connue au décodeur. Afin de mener les calculs à leurs termes, nous réintroduisons les hypothèses simplificatrices du CT classique opérant à haute résolution. Les schémas considérés nécéssitent donc que ni la transformation (KLT ou LDU), ni les paramètres de l'allocation de bits ne soient transmis au décodeur. Nous supposons par conséquent que ces schémas sont basés sur un estimé $\widehat{R} = R_{\underline{x}\underline{x}} + \Delta R$ de la matrice de covariance inconnue $R_{\underline{x}\underline{x}}$. $R_{\underline{x}\underline{x}}$ correspond à un processus vectoriel Gaussien $\underline{x}$ (éventuellement localement) stationnaire. $\widehat{R}$ est l'estimé correspondant, disponible au codeur et au décodeur. Dans ce cas, le processus de codage utilise une transformation $\widehat{T} = T + \Delta T$ (où T est la transformation calculée au moyen de $R_{\underline{x}\underline{x}}$), et la distorsion est proportionnelle aux variances ${\sigma'}^2_{y_i}$ des signaux transformés au moyen de $\widehat{T}$ au lieu de T. De plus, les bits $\widehat{r}_i$ sont attribués au moyen des estimés des variances disponibles au décodeur, notées $(\widehat{T}\widehat{R}\widehat{T})_{ii}$, où $(.)_{ii}$ dénote le $i$ème élément diagonal de $(.)$. Les résultats de cette allocation de bits "en ligne" sont par conséquent

$$\widehat{r}_i = r + \frac{1}{2}\log_2 \frac{(\widehat{T}\widehat{R}\widehat{T}^T)_{ii}}{(\prod_{i=1}^{N}(\widehat{T}\widehat{R}\widehat{T}^T)_{ii})^{\frac{1}{N}}}. \tag{10.5}$$

Nous obtenons alors la mesure de distorsion suivante, pour un schéma basé sur $\widehat{R}$, utilisant une transformation $\widehat{T}$:

$$\mathrm{E}\,||\underline{\tilde{y}}||^2_{\widehat{T}} = \mathrm{E}\sum_{i=1}^{N} c2^{-2\widehat{r}_i}{\sigma'}^2_{y_i} = \mathrm{E}\sum_{i=1}^{N} c2^{-2[r+\frac{1}{2}\log_2 \frac{(\widehat{T}\widehat{R}\widehat{T}^T)_{ii}}{(\prod_{i=1}^{N}(\widehat{T}\widehat{R}\widehat{T}^T)_{ii})^{\frac{1}{N}}}]}{\sigma'}^2_{y_i} \quad , \tag{10.6}$$

où l'espérance E correspond aux cas où ΔR est non déterministe.

Le but de ce travail est de fournir les expressions des distorsions correspondantes pour la KLT et la LDU, et de les comparer. Ces calculs sont faits dans trois cas.

Dans un premier cas, ΔR est créé par le bruit de quantification: le schéma de codage est basé sur les statistiques des données quantifiées ($\widehat{R} = R_{\underline{x}^q\underline{x}^q}$). Dans un second cas, ΔR correspond à un bruit d'estimation:

le système est basé sur un estimé de $R_{\underline{x}\underline{x}}$ construit au moyen de K vecteurs: $\widehat{R} = \frac{1}{K}\sum_{i=1}^{K} \underline{x}_i \underline{x}_i^T$. Finalement, les deux bruits sont traités ensemble: $\widehat{R} = \frac{1}{K}\sum_{i=1}^{K} \underline{x}_i^q \underline{x}_i^{qT}$.

Calculant dans chacun de ces trois cas la distorsion obtenue pour une transformation Identité (absence de transformation), puis pour la KLT et la LDU, nous obtenons des expressions analytiques pour le gain de codage.

Dans le cas où seul le bruit d'estimation est pris en compte, on montre quel les gains de codage sont les mêmes pour la KLT et la LDU. Dans le cas où les bruit de quantification et d'estimation sont pris en compte conjointement, on obtient pour la LDU

$$G_{\widehat{L}',K,q} = \frac{\mathrm{E}\,\|\underline{\tilde{x}}\|^2_{I,K,q}}{\mathrm{E}\,\|\underline{\tilde{y}}\|^2_{\widehat{L}',K,q}} \approx G_{TC}^{(0)} \frac{\left(\det(I + \sigma_q^2(\,\mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{-1})\right)^{1/N}}{\left(\det(I + \sigma_q^2(R_{\underline{x}\underline{x}})^{-1})\right)^{1/N}}$$
$$\times \frac{\left[1 + \frac{1}{K}\left[1 - \frac{1}{N^2}\,\mathrm{tr}\,\{R_{\underline{x}^q\underline{x}^q}(\,\mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{q-1} R_{\underline{x}^q\underline{x}^q}(\,\mathrm{diag}\,\{R_{\underline{x}\underline{x}}\})^{q-1}\}\right] - \frac{\sigma_q^2}{N}\,\mathrm{tr}\,\{(\,\mathrm{diag}\,R_{\underline{x}^q\underline{x}^q})^{-1}\}\right]}{\left[1 + \frac{N-1}{K}\left[\frac{1}{2} + \frac{1}{N}\right] - \frac{\sigma_q^2}{N}\,\mathrm{tr}\,\{(L' R_{\underline{x}^q\underline{x}^q} L'^T)^{-1}\}\right]}. \tag{10.7}$$

Cette expression (basée sur un calcul des perturbations au premier ordre, et supposant le nombre de vecteurs K suffisamment grand) permet de décrire quantitativement les influences respectives du bruit de quantification (termes en ^{q} et σ_q^2), du bruit d'estimation (termes en K), et leur influence conjointe (termes croisés). Comparant à l'expression correspondante pour la KLT, on montre ainsi que le bruit de quantification lié à l'utilisation de vecteurs de références quantifiés décroît à bas débit les performances de la LDU relativement à celles de la KLT. Les calculs théoriques de ce chapitre sont ensuite validés par des simulations numériques. Ces résultats sont présentés dans [57, 58, 59].

Notre but initial de présenter une analyse précise et complète de systèmes de CT adaptatifs "en ligne" nous a semblé toutefois partiellement inachevé à ce stade. En effet, les hypothèses simplificatrices retenues pour les calculs (principalement le mécanisme d'allocation optimale de bit) peuvent ne pas être réalistes pour des systèmes concrets. Ceci nous a mené aux développements du chapitre 4.

10.2.4 Analyse Débit-Distorsion de Schémas Concrets de Codage par Transformée Adaptatifs "en ligne"

Nous étudions dans ce chapitre, trois schémas concrets de CT "en ligne" basés sur la KLT et la LDU. Dans ces algorithmes, les quantificateurs scalaires sont de type ECUQ, et les pas de quantification sont les mêmes pour chaque composante transformée. Les transformations sont calculées sur la base des estimés des matrices de covariance obtenues à partir des données précédemment quantifiées.

Dans un premier temps, des algorithmes à pas de quantification constant (relativement au temps) sont implémentés. Ce cas présente un intérêt pour des sources stationnaires; autrement, de tels algorithmes peuvent occasionner des variations inacceptables de débit. Pour ces algorithmes, la question est de savoir si les transformations vont converger ou non vers les transformations "optimales" (*i.e.* les transformations calculées avec une connaissance parfaite des statistiques de la source). Nous montrons empiriquement que c'est le cas pour la LDU comme pour la KLT, même à très bas débit.

Dans un deuxième temps, nous proposons d'évaluer analytiquement le comportement de deux algorithmes à pas de quantification adaptatifs, permettant au système de produire un débit relativement constant. Pour ce problème, nous supposons que la source est un processus vectoriel (de dimension N) stationnaire (éventuellement par morceaux), de matrice de covariance R, inconnue du décodeur. La question est de savoir si le système CT adaptatif "en ligne" va converger vers un système créé avec la connaissance de R, *i.e.* vers un système produisant un débit r_0, et une distorsion $D_0 = c2^{-2r_0}(\det R)^{\frac{1}{N}}$. La procédure d'adaptation du pas de quantification est simple, et similaire à celle utilisée classiquement dans des schémas de quantification scalaire adaptative. Ces algorithmes sont les suivants:

Algorithme $[\mathit{1}]$:

- Initialisation: $K = N$.
- Etape 1: Un estimé de la matrice covariance $\widehat{R}_K = \frac{1}{K}\sum_{i=1}^{K} \underline{x}_i^q \underline{x}^{q\,T}_{\;i}$ est disponible au codeur et au décodeur.
- Etape 2: Une transformation $\widehat{T}_K$ est calculée de telle sorte que $\widehat{T}_K \widehat{R}_K \widehat{T}_K^T$ soit diagonale; $\widehat{T}_K$ est soit une KLT, soit une factorisation LDU de $\widehat{R}_K$. Un pas de quantification $\widehat{\Delta}_K^{[\mathit{1}]}$ est calculé par

$$\widehat{\Delta}_K^{[\mathit{1}]} = \sqrt{2\pi e}2^{-r_0} \det(\widehat{T}_K \widehat{R}_K \widehat{T}_K^T)^{\frac{1}{2N}}. \tag{10.8}$$

- Etape 3: Ces paramètres sont utilisés pour transformer et quantifier le $(K+1)$ème vecteur par: $\underline{y}_{K+1}^q = [\widehat{V}_K \underline{x}_{N+1}]_{\widehat{\Delta}_K^{[\mathit{1}]}}$ dans le cas unitaire, et par $\underline{y}_{K+1}^q = [\underline{x}_{K+1} - \widehat{\overline{L}} \underline{x}_{K+1}^q]_{\widehat{\Delta}_K^{[\mathit{1}]}}$ dans le cas causal ($[.]_\Delta$ dénote la quantification uniforme de pas Δ). L'espérance de la distorsion pour le $(K+1)$ème vecteur est alors $D^{[\mathit{1}]}(K+1) = \mathrm{E}\, \widehat{\Delta}_K^{[\mathit{1}]^2}/12$.
- Etape 4: Retour à l'Etape 1: le décodeur calcule un estimé de la matrice de covariance $\widehat{R}_{K+1} = \frac{1}{K+1}(\sum_{i=1}^{N} \underline{x}_i^q \underline{x}_i^{qT} + \underline{x}_{K+1}^q \underline{x}_{K+1}^{qT})$, à partir duquel $\widehat{T}_{K+1}$ et $\widehat{\Delta}_{K+1}$ peuvent être calculés, utilisés pour coder le $(N+2)$ème vecteur, etc...

Algorithme $[\mathit{2}]$:

Une amélioration simple à l'algorithme précédent peut être apportée en utilisant des résultats concernant la quantification uniforme de sources Gaussiennes. Pour des vecteurs Gaussiens $\underline{y}_i$, quantifiés avec le même pas de quantification Δ, on montre que $\mathrm{E}\, \underline{y}_i^q \underline{y}_i^{qT} = R_{\underline{y}^q \underline{y}^q} = R + \frac{\Delta^2}{12} I + B,$ où $B \to 0$ élément par élément quand $\Delta \to 0$. Dans l'algorithme $[\mathit{2}]$, si le pas de quantification converge vers un certain pas (suffisamment petit) $\Delta_\infty(T)$, l'estimé de la matrice de covariance converge alors vers une matrice proche de $R + \frac{\Delta_\infty^2(T)}{12} I$. Les évaluations numériques de la première partie de ce chapitre ont suggéré la convergence des estimés $\widehat{R}$ vers $R + \frac{\Delta^2}{12} I$, même pour des pas de quantification de l'ordre de l'écart-type des sources scalaires. Par conséquent, un estimé plus précis de $\widehat{R}$ peut être obtenu en soustrayant $\frac{\widehat{\Delta}_K^2}{12} I$ à l'estimé actuel après un certain nombre N_1 de vecteurs codés/décodés. Cette correction sur l'estimation des moments de second ordre d'une source au moyen de sa version quantifiée est parfois appelée "correction de Sheppard". A part cette différence sur l'estimé $\widehat{R}$ intervenant après N_1 vecteurs, les étapes de l'agorithme $[\mathit{2}]$ sont les mêmes que celles de l'agorithme $[\mathit{1}]$.

Nous modélisons ensuite l'espérance de la distorsion obtenue pour chaque algorithme, pour un nombre donné K de vecteurs décodés, et obtenons les expressions récursives suivantes

$$D^{[2]}_{(K+1)} \approx D_0 \left[1 + \frac{1}{K}\left(\frac{1}{N} - N\right) + \operatorname{tr}\{R^{-1}\} \left(\frac{1}{N} \left[\frac{1}{K} \left(\sum_{i=N+1}^{N_1} D^{[1]}_{(i)} + \sum_{i=N_1+1}^{K} D^{[2]}_{(i)} \right) - D^{[2]}_{(K)} \right] \right) \right], \qquad (10.9)$$

$$D^{[1]}_{(K+1)} \approx D_0 \left[1 + \frac{1}{K}\left(\frac{1}{N} - N\right) + \frac{\operatorname{tr}\{R^{-1}\}}{KN} \left(\sum_{i=N+1}^{K} D^{[1]}_{(i)} \right) \right]. \qquad (10.10)$$

Par conséquent, nous montrons que si D_0 est la distorsion cible, choisie pour un débit r_0 et une source de covariance R, l'algorithme utilisant la correction de Sheppard converge vers le point (r_0, D_0, Δ_0) choisi. Si cette correction n'est pas appliquée, le système CT adaptatif converge vers une distorsion plus grande $D_0 + \delta D_0$, et un débit plus petit $r_0 - \delta r_0$:

$$\begin{cases} D^{[2]}_{\infty} \approx D_0 \\ \Delta^{[2]}_{\infty} \approx \Delta_0 \\ r^{[2]} \approx r_0 \end{cases} \quad , \text{ et } \quad \begin{cases} D^{[1]}_{\infty} \approx D_0 + \overbrace{D_0^2 \frac{\operatorname{tr} R^{-1}}{N}}^{\delta D_0} \\ \Delta^{[1]}_{\infty} \approx \Delta_0 (1 + \Delta_0 \frac{\operatorname{tr} R^{-1}}{24N}) \\ r^{[1]} \approx r_0 - \underbrace{\frac{D_0}{2N \ln 2} \operatorname{tr}\{R^{-1}\}}_{\delta r_0}. \end{cases}$$

Les résultats numériques confirment les assertions théoriques. En particulier, le système utilisant la correction de Sheppard converge vers le point (r_0, D_0, Δ_0) cible choisi au décodeur, bien que le décodeur n'ait *a priori* aucune connaissance du pas de quantification à utiliser, ni des statistiques de la source à coder. Ces résultats sont présentés dans [60].

La transformation causale étudiée dans ces premiers chapitres possède des propriétés de décorrélation optimales. Comme la KLT toutefois, la transformation causale telle que décrite jusqu'à présent ne prend en compte que les corrélations à l'intérieur de chaque bloc (dépendances *spatiales*). Pour des sources vectorielles dont les échantillons vectoriels ne sont pas indépendants, une efficacité de codage supérieure peut être obtenue en prenant en compte les dépendances *temporelles*. La description et l'étude de la transformation causale dans ce cadre est l'objet du chapitre suivant.

10.2.5 Prédiction MIMO (Multiple Input Multiple Output, multi-entrées multi-sorties) Généralisée

Nous montrons d'abord dans ce chapitre comment la transformation causale LDU peut être étendue au filtrage matriciel. Nous supposons que les échantillons $\underline{x}_k$ (de tailles $M \times 1$) sont collectés dans un "super-vecteur" $\underline{X}_k = [\underline{x}_0^T \ \underline{x}_1^T \ \cdots \underline{x}_k^T]^T$, et considérons le cas limite où $k \to \infty$: dans ce cas, la matrice L peut être décrite par une matrice $L(z)$ (de taille $M \times M$) :

- Pour $\underline{X}_k$ décrit précédemment, la matrice $L(z)$ correspond à la prédiction MIMO classique. Pour $M = 2$,

$$L(z) = \begin{bmatrix} L_{11}(z) & L_{12}(z) \\ L_{21}(z) & L_{22}(z) \end{bmatrix} = \sum_{k=0}^{\infty} L_k z^{-k} \text{ avec } L_0 = \begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix},$$

  afin de conserver la structure temporellement causale.

- En organisant différemment les échantillons au sein de $\underline{X}_k$, nous obtenons la prédiction MIMO triangulaire

$$L(z) = \begin{bmatrix} 1 & 0 \\ 0 & L_{22}(z) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ W_{21}(z) & 1 \end{bmatrix} \begin{bmatrix} L_{11}(z) & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} L_{11}(z) & 0 \\ L_{22}W_{21}L_{11} & L_{22}(z) \end{bmatrix}, \tag{10.11}$$

où W_{21} est un filtre de Wiener. Comparant la prédiction MIMO classique à sa contrepartie triangulaire, les degrés de liberté de L_{12} sont transférées à la partie anticausale de L_{21}. Les filtres diagonaux sont des filtres scalaires de prédictions (intrasignaux), et les filtres non diagnaux sont des filtres de Wiener réalisant une décorrélation intersignaux. Nous montrons que la prédiction MIMO classique et la prédiction MIMO triangulaire sont deux cas particuliers d'une infinité de manières de décorréler les signaux vectoriels via Gram-Schmidt. Ces différentes approches sont caractérisées par le degré d'anticausalité dédié aux filtres non diagonaux, et peuvent être vues comme des prédictions MIMO classiques appliquées à des signaux vectoriels $\underline{x}'_k = [x_{1,k} \quad x_{2,k+d_1} \ldots x_{M,k+d_1+\ldots+d_{M-1}}]^T$, où les d_i sont des délais. Nous montrons alors que la prédiction MIMO triangulaire est aussi "causale", mais dans un sens plus large:

- elle correspond au cas extrême où les délais $d_i \to \infty$, $i = 1, ..., M-1$,
- pour la matrice de prédiction triangulaire,
  - la notion de causalité reste inchangé pour les prédicteurs diagonaux (SISO, Single-Input Single-Output),
  - les filtres non diagonaux sont des filtres de Wiener entre des signaux scalaires,
  - la causalité entre les canaux correspond à l'*ordre* dans les signaux scalaires sont décorrélés.

Par conséquent, certains signaux peuvent être codés en utilisant les versions codées/décodées des "précédents" signaux. Ainsi, la prédiction MIMO triangulaire, cas particulier d'une prédiction MIMO généralisée, apparaît comme une généralisation au cas vectoriel de la technique (A)DPCM.

Une question intéressante est alors l'étude des gains de codage pour ces approches décorrélatrices causales. En considérant des vecteurs de taille infinie, on obtient des expressions fréquentielles pour le gain de codage. Pour les mêmes raisons que dans le cas de la LDU, une implémentation réaliste de ces systèmes devrait être faite en boucle fermée. Par conséquent, l'analyse théorique proposée comporte encore deux étapes.

Dans un premier temps, on néglige le fait que la prédiction soit faite en utilisant des données quantifiées (hypothèse de résolution infinie). Dans ce cas, on montre que toutes les approches de la prédiction MIMO généralisée (notamment classique et triangulaire) sont équivalentes, et que le gain de codage associé est

$$G^{(0)} = \left(\frac{\prod_{i=1}^{M} \sigma_{x_i}^2}{e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln[\det(S_{\underline{x}\underline{x}}(f))]\, df}} \right)^{\frac{1}{M}}, \tag{10.12}$$

où $S_{\underline{x}\underline{x}}(f)$ est la matrice de densité spectrale du processus $\underline{x}$.

Pour des systèmes en boucle fermée utilisant des données quantifiées à haute résolution dans un deuxième temps, on montre que le gain de codage est

$$G^{(1)}(L) \approx G^{(0)} \left[1 + \frac{\sigma_q^2}{M} \left(- \int_{-\frac{1}{2}}^{\frac{1}{2}} \text{tr} \left(S_{\underline{x}\underline{x}}^{-1}(f) \right) df + \sum_{i=1}^{M} \frac{1}{\sigma_{y_i}^2} \right) \right], \tag{10.13}$$

où σ_q^2 est la variance de l'erreur de quantification dans le cas idéal (résolution infinie), et où les $\sigma_{y_i}^2$ sont les variances de prédiction optimales.

Ainsi, pour une résolution infinie, toutes les approches décorrélation sont équivalentes ($G^{(0)}$), alors que pour une haute résolution, maximiser $G^{(1)}(L)$ équivaut à maximiser $\sum_{i=1}^{M} \frac{1}{\sigma_{y_i}^2}$. Nous proposons un théorème pour ce problème, qui montre que l'ordre optimal dans la décorrélation pour le prédicteur triangulaire est de décorréler les signaux par ordre de variance décroissante.

Le cas de filtres à réponses impulsionnelles finies (RIFs), ainsi que celui d'une décorrélation opérée dans le domaine fréquentiel sont ensuite abordés. Finalement, une application directe de ces résultats est proposée pour le codage de la parole large bande. Ces résultats, ainsi qu'une démonstration audio ont été présentés à [61].

10.2.6 Origines des Précédents Résultats: Structures Analyse par Synthèse

Les résultats du travail présenté dans cette thèse trouvent leur origine dans le projet RNRT *COBASCA*[4], dont le but était de fournir des algorithmes de codage conjoint source-canal pour des signaux audios large bande ($[50Hz - 7kHz]$) dans le contexte d' UMTS. Nous avons pour cela suivis deux axes de recherche. Le premier concerne l'optimisation conjointe des prédicteurs linéaires à court et long terme pour des signaux de parole; ce sujet sort néanmoins du cadre de cette thèse (les résultats associés sont reportés dans [101, 102]). Il nous a cependant semblé intéressant de fournir un descriptif du second axe de recherche, parce qu'il montre comment des techniques de codage existantes, des contraintes industrielles, et des objectifs scientifiques ont conjointement mené à l'ensemble de techniques causales de codage présentées dans cette thèse.

[4] *CO*dage en *B*ande élargie avec partage *A*daptatif du débit entre *S*ource et *CA*nal pour réseaux cellulaires de deuxième et troisième générations (UMTS), http://www.telecom.gouv.fr/rnrt/pcobasca.html.

10.3 Deuxième Partie: Codage Causal sans Pertes

La seconde partie de cette thèse présente et analyse des techniques de codage causales sans pertes basées sur les approches décorrélantes (de type LDU et MIMO généralisée) décrites dans la première partie. Nous présentons d'abord les principales problématiques de cette partie. Sommairement, les structures de codage étudiées mettent en oeuvre des transformations non linéaires, les transformations d'entiers à entiers, et abordent le problème du codage multirésolution.

Transformations d'Entiers à Entiers

Les schémas de codage sans pertes peuvent exister comme des codeurs à part entière (codeurs "entropiques", par exemple de type Huffman), ou bien être inclus dans la structure de codeurs avec pertes, afin d'en améliorer les performances. Considérons dans ce cas le schéma de la figure 10.2, qui utilise une transformation transformation T.

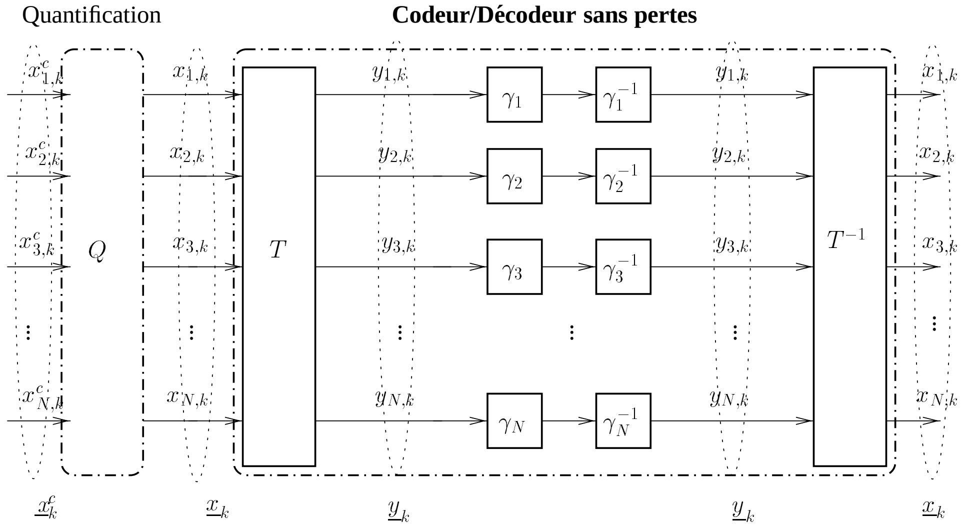


Figure 10.2: Schéma de codage sans pertes emboîté dans un codeur avec pertes.

Dans un premier temps, une source vectorielle à très haute résolution $\underline{x}^c$ (amplitude continue) est quantifiée au moyen d'un codeur avec pertes représenté par le bloc Q (Q peut représenter la discrétisation en amplitude réalisée par des quantificateurs indépendants de type PCM, des structures de type ADPCM, des codecs MPEG, etc...). Une fois cette discrétisation réalisée, le problème est de transmettre efficacement la source discrète $\underline{x}$ ou, en d'autres termes, de minimiser le débit associé à la représentation de cette source. Une méthode de codage entropique optimale est un codage entropique vectoriel, qui assigne des mots de code à des vecteurs. Cependant, cette méthode requiert de calculer la distribution de probabilité conjointe des vecteurs sources; elle est par conséquent complexe, mal adaptée à des signaux présentant des corrélations à long terme [5], et peu utilisée en pratique. Dans ce cas, on préfère coder l'ensemble des flux x_i par N codeurs entropiques scalaires (indépendants) γ_i [6]. Bien sûr, ce type de codage entropique scalaire est sub-optimal parce que les sources x_i ne sont pas indépendantes, et que les flux séparément transmis comportent des redondances; il est néanmoins largement moins complexe. Une façon de pallier à cette suboptimalité est d'appliquer, après l'étage de quantification et avant le codage entropique, une transformation T sans pertes, ou réversible, qui rend ces flux indépendants (ou au moins décorrélés). Les redondances intersignaux étant réduites, le débit total nécéssaire à la représentation des sources ainsi transformées s'en trouve réduit également. La transformation T s'appuyant sur un ensemble discret, et produisant un autre ensemble discret, elle est non linéaire, et appelée transformation d'entiers à entiers.

Pour résumer, une approche de codage utlisant une transformation d'entiers à entiers sépare la procédure de codage entropique en deux étapes: premièrement la transformation inversible est appliquée à chaque bloc quantifié dans un but de décorrélation; deuxièmement, les coefficients transformés sont indépendamment codés, ce qui assure une complexité totale relativement faible. Du signal vectoriel $\underline{x}$ on passe au signal $\underline{y}$, à partir duquel le décodeur peut retrouver exactement le signal $\underline{x}$. Cette approche sera aussi appelée codage sans pertes "monoétage" ou "monorésolution".

Pour une transformation T et une source $\underline{x}$ données, nous allons considérer deux scénarios: le scénario 1, où T est utilisée, et le scénario 2, où elle ne l'est pas. Dans les deux cas, la structure de codage utilise N codeurs entropiques scalaires. Les questions suivantes se posent alors: quelle est la réduction maximale de débit que le schéma 2 peut opérer relativement au schéma 1, et quelle serait alors la transformation correspondante ? Deuxièmement, quelle est la réduction de débit réellement opérée par des transformations T concrètes ? Dans le chapitre 6, ces transformations T concrètes sont basées sur les implémentations entiers à entiers de la KLT et de la LDU. Dans le chapitre 8, T est basée sur les approches de MIMO généralisée.

Codage sans Pertes Multirésolution

Parallèlement à cette approche monorésolution, une approche de codage sans pertes différente consiste à coder avec pertes la source $\underline{x}$ dans un premier temps, produisant par là un premier flux de N signaux scalaires "basse résolution" y_i^q. Dans un second temps, le signal d'erreur est encodé séparément, ce qui donne la structure à deux étages de la figure 10.3.

L'avantage de ce type de schémas est qu'une version approximative de la source peut être disponible rapidemment, indépendamment du signal d'erreur (*e.g.* dans le cas où la capacité du lien de transmission varie, sur internet par exemple). Le signal original peut être reconstruit ultérieurement en ajoutant le signal d'erreur. Si l'on suppose que $\{Q\}$ est composé de quantificateurs scalaires, le débit du signal basse résolution $\underline{x}^q$ de $\underline{x}$ peut être contrôlé simplement par les pas de quantification correspondants. Ceci permet d'obtenir un signal bas débit, au coût d'une certaine distorsion. Ce type de schémas est utilisé dans

[5] Pour des sources vectorielles avec mémoire, le problème est plus aigü puisqu'elle nécéssite d'estimer la probabilité conjointe de vecteurs *de vecteurs*.

[6] Par exemple, des codes populaires en audio sont les codes de Huffman et de Golomb-Rice.

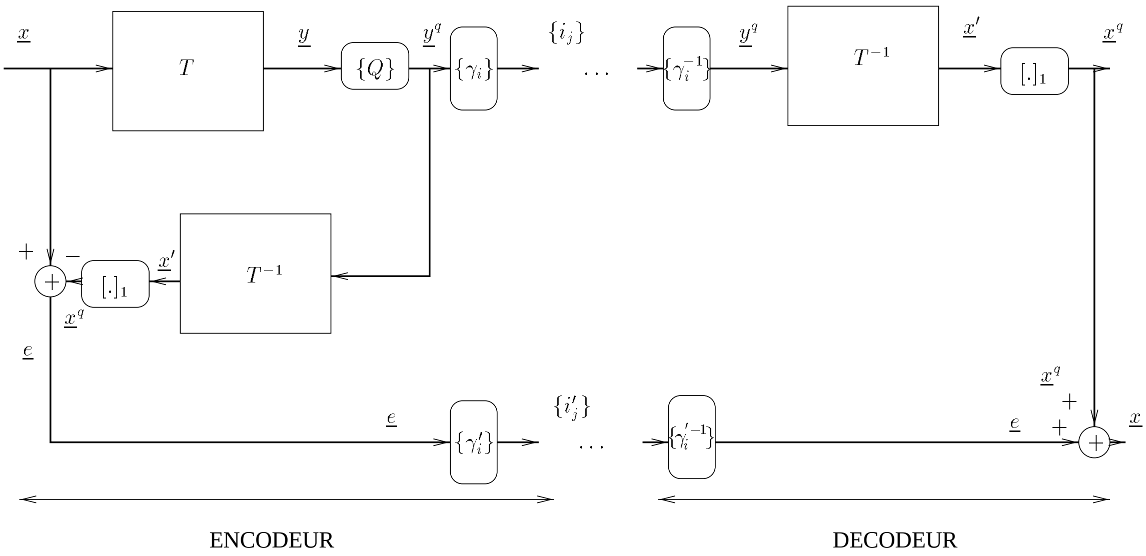


Figure 10.3: Schéma classique du CT sans pertes à deux niveaux de résolution. $\{Q\}$ dénote des quantificateurs scalaires , $\{\gamma_i\}$ et $\{\gamma_i'\}$ des codeurs entropiques scalaires, et $[.]_1$ et des opérateurs d'arrondi.

des codeurs de signaux audio sans pertes [21, 24], et d'images [105, 106]. Par conséquent, une comparaison de l'efficacité de compaction entre les transformations orthogonales traditionnelles et la transformation causale semble intéressante. Par ailleurs, une question d'intérêt est celle de savoir si un schéma de codage multirésolution sans pertes est sous optimal par rapport à l'approche monorésolution décrite plus haut. Ces questions sont traitées dans le chapitre 7 pour des approches à deux étages basées sur la LDU ou sur les transformations orthogonales, et dans le chapitre 8 pour des structures à 2 et M étages basées sur des prédicteurs MIMO.

10.3.1 Codage par Transformée sans Pertes: Cas Causal, Unitaire, et Étude de Systèmes "en ligne"

Le chapitre 6 adresse le problème du CT sans pertes monoétage. Dans le cas où T de la figure 10.2 est basé sur des matrices décorrélantes de type KLT ou LDU, la relation de $\underline{x}^c$ à $\underline{y}$ est similaire à celle existant dans le CT classique, sauf que les opérations de quantification et de transformations apparaissent *en ordre inverse*; de plus, les signaux transformés doivent être à amplitude discrète puisqu'ils sont par la suite codés entropiquement. La question de savoir si les transformations d'entiers à entiers sont, d'un point de vue débit/distorsion, aussi efficaces que leurs contreparties linéaires a été adressée récemment [41]. Supposons que l'étage de quantification Q soit composé de N quantificateurs uniformes de même pas Δ, et considérons les schémas de codage suivants:

- (1) quantification scalaire des x_i^c suivie de N codeurs entropiques scalaires,

- (2) quantification scalaire des x_i^c suivie d'une transformation d'entiers à entiers et de N codeurs entropiques scalaires (Q et CT sans pertes),
- (3) transformation décorrélante linéaire suivie d'une quantification scalaire, suivie de N codeurs entropiques scalaires (CT),
- (4) quantification scalaire des x_i^c suivie d'un codage entropique vectoriel.

Les résultats de [41] montrent que, pour des vecteurs Gaussiens i.i.d., les performances des schémas (2), (3), et (4) sont équivalentes dans la limite de petits pas de quantification Δ. Cette analyse revient à négliger la contrainte d'entiers à entiers sur les transformations du cas (2). En effet, ces transformations doivent produire des coefficients discrets; elles ne sont pas linéaires et ne peuvent qu'approximer leur contrepartie linéaire. Le but de ce chapitre est d'évaluer la sous-optimalité liée à ces non-linéarités. Le critère choisi pour cette évaluation est un *gain de codage sans pertes*, défini comme la réduction de débit opérée par le schéma (2) par rapport au schéma (1) (en bit par échantillon).

Nous montrons d'abord que les gains des schémas (3) et (4) représentent un limite supérieure au gain du schéma (2). Le débit minimum nécéssaire au codage sans pertes de signaux sans mémoire est l'entropie discrète. Pour des signaux Gaussiens, nous utilisons la relation de Rényi [38]

$$H(x_i^q) \approx \frac{1}{2}\log_2 2\pi e\sigma_{x_i}^2 - \log_2 \Delta_i. \tag{10.14}$$

Le gain de codage sans pertes maximal, obtenu pour un codage vectoriel et pour une source sans mémoire, est alors donné par

$$G_{max} = \frac{1}{2N}\log_2 \frac{\det \text{diag}\{R_{\underline{xx}}\}}{\det R_{\underline{xx}}} = \frac{1}{N}\sum_{i=2}^{N} \underbrace{I(x_i;\underline{x}_{1:i-1})}_{\text{Information mutuelle}} \tag{10.15}$$

Cette expression montre que pour des signaux Gaussiens, la réduction de débit opérée par le schéma (4) sur le schéma (1) correspond à la moyenne des informations mutuelles entre chaque nouvelle variable x_i^q du vecteur $\underline{x}^q$ et les variables précédemment codées $x_{1:i-1}^q$. Elle permet aussi de donner une interprétation du gain de codage traditionnel en termes de l'information mutuelle.

Nous comparons ensuite à cette limite les gains réellement opérés par les implémentations sans pertes de la LDU et de la KLT, pour un niveau de distorsion fixé (Q fixé). Nous montrons d'abord que l'implémentation entiers à entiers de la LDU peut être obtenue très simplement grâce à sa structure triangulaire. Le gain sans pertes associé à la LDU est donné par

$$G_{L_{int}^q} \approx G_{max} - \frac{1}{2N\ln 2}\left[\text{tr}\{R_{\underline{xx}}^{-1}D\} - \frac{\Delta_1^2}{12\sigma_{x_1}^2}\right]. \tag{10.16}$$

Une conséquence intéressante de ce résultat est que la version la plus grossièrement quantifiée ($\frac{\Delta_i}{\sigma_{x_i}^2}$) doit être placée en première position pour maximiser le gain (minimiser le débit).

En ce qui concerne la KLT, nous suivons la factorisation donnée par Goyal pour obtenir la transformation d'entiers à entiers associée. Dans le cas $N = 2$ par exemple, si V^q est une KLT de la source à coder, on la

factorise comme $V^q = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = V_1^q\, V_2^q\, V_3^q; \quad V_1^q = \begin{bmatrix} 1 & \frac{a-1}{c} \\ 0 & 1 \end{bmatrix}, V_2^q = \begin{bmatrix} 1 & 0 \\ c & 1 \end{bmatrix}, V_3^q = \begin{bmatrix} 1 & \frac{d-1}{c} \\ 0 & 1 \end{bmatrix}$.
La transformation recherchée V_{int}^q est alors obtenue en intercalant des opérations de quantification Δ_i après chaque matrice V_i^q. Analysant les effets liés aux non linéarités introduites par les Δ_i sous l'hypothèse de haute résolution, on montre que le gain de codage est strictement inférieur dans le cas de la KLT par rapport au cas causal: $G_{V_{int}^q} < G_{L_{int}^q}$. L'approche causale mène à une réduction de débit plus importante que l'approche unitaire, ce qui est une conséquence de sa structure triangulaire.

Finalement, l'adaptativité de systèmes sans pertes monorésolution "en ligne" est étudiée: nous considérons des systèmes pour lesquels les transformations d'entiers à entiers sont calculées sur la base des données précédemment reçues au décodeur uniquement, c'est à dire au moyen d'un estimé de la matrice de covariance de type $\widehat{R} = \frac{1}{K}\sum_{i=1}^{K} \underline{x}_i^q \underline{x}_i^{qT}$. Dans ce cas, les transformations convergent vers les transformations "optimales" (basées sur R) seulement quand le nombre de vecteurs K tend vers l'infini. La question est ici: quelle est la réduction de débit moyenne $G_{\widehat{T}_{int}}(K)$ apportée par une transformation $\widehat{T}_{int}$ calculée avec $\widehat{R}$ basé sur K vecteurs ? Nous calculons pour ce problème un modèle statistique de vecteurs Gaussiens i.i.d. :

$$\mathrm{E}\,\mathrm{vec}(\Delta R)\,(\mathrm{vec}(\Delta R))^T \approx \frac{2}{K} R_{\underline{x}^q\underline{x}^q} \otimes R_{\underline{x}^q\underline{x}^q}\,, \tag{10.17}$$

et obtenons, pour K suffisamment grand, des gains en fonction de K donnés par

$$\begin{aligned} G_{\widehat{L}_{int}^q}(K) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} H(x_i^q) - \mathrm{E}\,H(y_i^q, K) \\ &\approx G_{L_{int}^q} - \tfrac{N-1}{4\ln 2K}. \end{aligned} \tag{10.18}$$

$$G_{\widehat{V}_{int}^q}(K) \approx G_{V_{int}^q} - \tfrac{N-1}{4\ln 2K}. \tag{10.19}$$

Les résultats analytiques de ce chapitre sont ensuite comparés à des résultats numériques obtenus en implémentant les systèmes étudiés. Ces travaux sont présentés dans [142].
Après l'analyse de systèmes monorésolution, la suite de cette partie se tourne vers des systèmes de codage sans pertes à deux niveaux de résolution basés sur la KLT et de la LDU.

10.3.2 Sur la Sous-Optimalité des Transformations Orthogonales pour le Codage par Transformée sans Pertes

Dans le chapitre 7, nous nous intéressons au schéma classique du CT sans pertes à deux niveaux de résolution de la figure 10.3. Pour un étage de quantification fixé et à haute résolution, nous analysons les débits r_{LR} et $\overline{r}$ nécessaires pour coder respectivement la version basse résolution et le signal d'erreur. Le débit total $r_{LR} + \overline{r}$ est comparé à celui obtenu pour le codeur monorésolution correspondant (basé soit sur la KLT, soit sur la LDU).
Pour la KLT, le schéma est celui de la figure 5.8; les statistiques des signaux d'erreurs pour une source Gaussienne ont été analysés dans [21]. Pour le schéma basé sur la LDU, le schéma correspondant est différent à cause de la structure de prédiction en boucle fermée, et est représenté par la figure ci dessous.

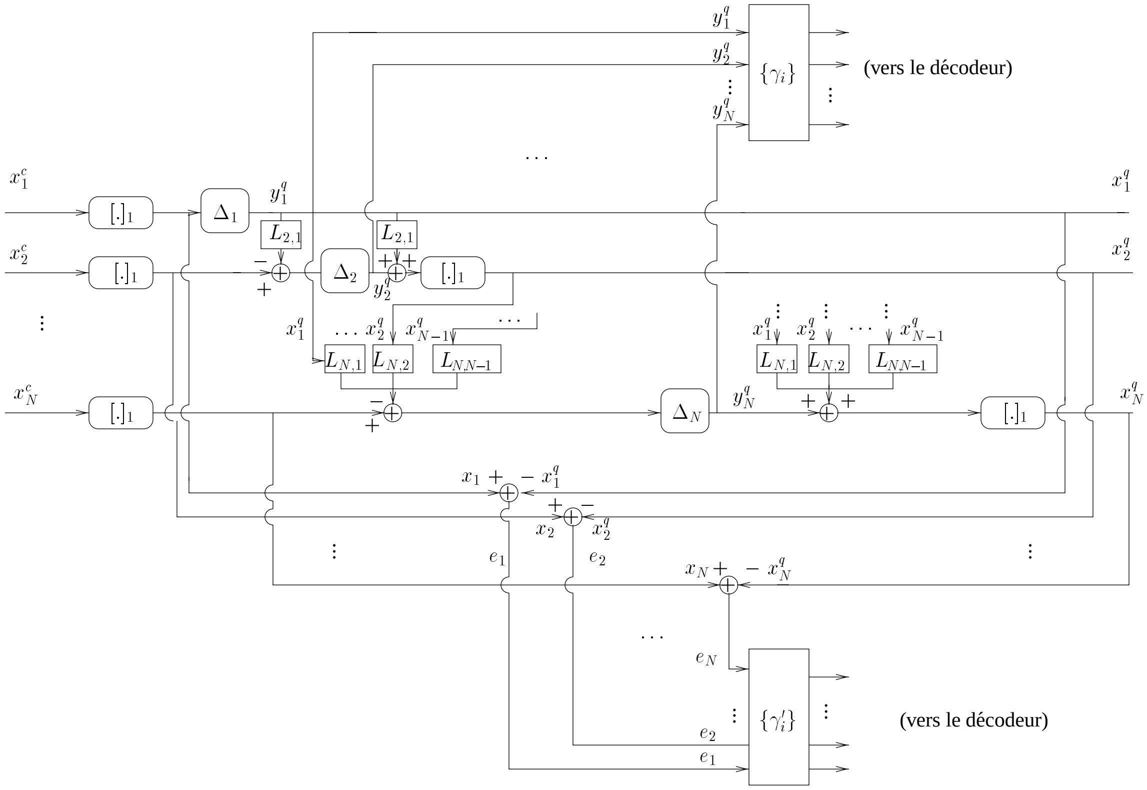


Figure 10.4: Encodeur du CT sans pertes à deux niveaux de résolution dans le cas causal.

Menant une analyse similaire à celle de [21], nous montrons qu'alors que les transformations orthogonales tendent à "gaussianniser" le signal d'erreur, la transformation causale les laisse approximativement uniformes [7]. Les probabilités qu'un échantillon du signal d'erreur e_i soit non nul sont données par

$$\text{Cas unitaire} : P(e_i \neq 0) = P(|e_i| \geq \frac{1}{2}) \approx 1 - erf(\sqrt{\frac{3}{2}}\frac{1}{\Delta}). \tag{10.20}$$

$$\text{Cas causal} : P(e_i \neq 0) = P(|e_i| \geq \frac{1}{2}) = 1 - \frac{1}{\Delta_i} \quad \forall \Delta_i \tag{10.21}$$

(dans le cas unitaire, les pas des quantificateurs sont tous égaux pour contrôler la distorsion totale; ce n'est pas nécéssaire dans le cas causal). Calculant l'entropie discrète associée à ces distributions de probabilités, nous obtenons dans le cas causal

$$\begin{array}{lcl} \text{Signal basse resolution} : r_{LR_{LDU}} & \approx & r_{LR_{KLT}} + \frac{\Delta^2}{24N\ln 2}\sum_{i=1}^{N}\left(\frac{1}{\lambda_i} - \frac{1}{\sigma^2_{y_i^0}}\right). \\ \text{Signal d'erreur} : \qquad \overline{r}_{LDU} & \approx & \overline{r}_{KLT} - \underbrace{\frac{1}{2}\log_2\frac{\pi e}{6}}_{\approx 0.25 \text{ bit } \forall\Delta}, \end{array} \tag{10.22}$$

[7] Elles sont strictement uniformes si les Δ_i sont impairs.

Nous montrons ainsi que dans le cas causal, le débit total est le même que dans le cas de codeur monorésolution, au terme $\frac{\Delta^2}{24N \ln 2} \sum_{i=1}^{N} \left(\frac{1}{\lambda_i} - \frac{1}{\sigma^2_{y_i^0}} \right)$ près, qui vient du fait que la prédiction est faite sur la base du signal basse résolution au lieu du signal original. Ce terme est néanmoins négligeable dans la grande majorité de situations pratiques de codage. Les débits pour les versions basse résolution sont donc sensiblement les mêmes dans les deux cas. A l'opposé, les débits des signaux d' erreur diffèrent dans le cas orthogonal et causal d'environ 0.25 bit/éch., qui correspond à la différence entre les entropies d'une variable aléatoire (v.a.) Gaussienne et d'une v.a. uniforme de mêmes variances. En conclusions, le débit total dans le cas causal est le même que dans le cas monorésolution, et il est environ 0.25 bit/éch. inférieur à celui obtenu pour la KLT.

De plus, nous soulignons que le schéma causal à deux niveaux de résolution présente des avantages pratiques très intéressants par rapport au schéma classique de la figure 10.2. Premièrement, ce schéma offre la possibilité de passer instantanément d'un schéma mono- à un schéma bi-résolution en fixant tous les pas de quantification à 1; ceci n'est pas possible dans le cas orthogonal (pour $\Delta = 1$ dans (10.20), $P(e_i \neq 0) \approx \frac{1}{12}$). Deuxièmement, un ou plusieurs canaux x_i peuvent être codés en monorésolution, et les autres en multirésolution. Finalement, le codage entropique du signal d'erreur devient très simple dans le cas causal, puisque la distribution étant dans certains cas exactement uniforme, transmettre la représentation binaire des échantillons est optimal. Ces travaux sont présentés dans [143].

Comme à la fin du chapitre 4, nous généralisons les résultats obtenus dans le début de cette deuxième partie en considérant des sources avec mémoire, et en appliquant la LDU à des vecteurs de vecteurs de taille arbitrairement grande, dans les contextes mono- et multi-résolution décrits ci-dessus.

10.3.3 Prédiction MIMO d'Entiers-à-Entiers Mono- et Multirésolution

Ce dernier chapitre traite du codage sans pertes "optimal" (minimisant le débit) pour les signaux vectoriels. La structure de codage etudiée tout d'abord est celle du schéma $(\mathit{2})$, ou encore de la figure 10.2, où la transformation T est une implémentation entiers à entiers d'un des prédicteurs MIMO décrits au chapitre 6. Premièrement, on cherche à exprimer la réduction de débit maximale relativement au schéma $(\mathit{1})$, pour une source $\underline{x}$ avec mémoire. Nous supposons que $\underline{x}$ est une version uniformément quantifiée de $\underline{x}^c$ avec des pas Δ_i, et que les échantillons de $\underline{x}$ sont collectés dans un vecteur $\underline{X}_k = [\underline{x}_1 \ldots \underline{x}_k]^T$. Par le théorème de codage sans bruit d'une source discrète, le débit minimum associé à la représentation de $\underline{x}$ est le débit entropique $r_0(\underline{x})$ de cette source,

$$r_0(\underline{x}) = \lim_{k \to \infty} \frac{1}{Nk} H(\underline{X}_k). \tag{10.23}$$

Exprimant ce débit pour une source Gaussienne, on obtient la réduction de débit maximale G_{Max} qu'il est possible d'opérer sur le schéma $(\mathit{1})$,

$$G_{Max} = r_{scal}(\underline{x}) - r_0(\underline{x}) \approx \frac{1}{2M} \log_2 \frac{\det \ \text{diag}\{R_{\underline{x}^c \underline{x}^c}\}}{e^{\int_{-1/2}^{1/2} \ln \det S_{\underline{x}^c \underline{x}^c}(f) df}}. \tag{10.24}$$

Cette réduction est possible avec un codage entropique vectoriel, asymptotiquement dans la longueur des vecteurs. La complexité de cette méthode la rend toutefois irréaliste pour une approche pratique. Nous montrons par la suite que ses performances peuvent néanmoins être approchées par des prédicteurs MIMO sans pertes.

Nous considérons ensuite les gains de codage sans pertes associés aux prédicteurs MIMO, et présentons les structures de codage associées. Nous montrons que pour toutes les approches décorrélantes du contexte de prédiction MIMO généralisée, l'implémentation entiers à entiers correspondante produit un gain

$$G_{L(z)} \approx \underbrace{\frac{1}{2M}\log_2 \frac{\det diag\{R_{\underline{x}^c\underline{x}^c}\}}{e^{\int_{-1/2}^{1/2} \ln\det S_{\underline{x}^c\underline{x}^c}\, df}}}_{G_{Max}} - \underbrace{\frac{\Delta^2}{24M\ln 2}\left(\int_{-1/2}^{1/2} \mathrm{tr}\ S^{-1}_{\underline{x}^c\underline{x}^c}(f)df\right)}_{Exces\ de\ debit:\ contrainte\ sans\ pertes}, \tag{10.25}$$

De même qu'au chapitre 6, les non-linéarités liées à la contrainte entiers à entiers se manifestent comme un excès de débit (diminution du gain) par rapport à la méthode de codage idéale. Elles deviennent néanmoins négligeables à haute résolution ($\Delta \to 0$).

Dans un second temps, nous généralisons le schéma TC à deux niveaux de résolution de la figure 10.4 au cas du filtrage. Les débits des signaux basses résolutions et des signaux d'erreurs sont contrôlés par des quantificateurs de pas égaux Δ_{Q_1}. Nous comparons alors le débit total produit par la structure obtenue au débit de la structure monorésolution présentée au début de ce chapitre. Nous obtenons, pour les débits des signaux basse résolution r_{LR}, et les signaux d'erreur $\overline{r}_L$

$$\begin{cases} r_{LR} \approx r_{one-shot}(\underline{y})(1+\underbrace{\frac{\Delta^2_{Q_1}}{24M\ln 2}\left[\int_{-1/2}^{1/2}\mathrm{tr}\, S^{-1}_{\underline{x}^c\underline{x}^c}(f)df - \sum_{i=1}^{M}\frac{1}{\sigma^2_{y_i^0}}\right]}_{Facteur\ d'excs\ de\ dbit}) \ -\underbrace{\log_2\Delta_{Q_1}}_{Reduction\ de\ debit}, \\ \overline{r}_L \approx \log_2 \Delta_{Q_1} \end{cases} \tag{10.26}$$

où M est la dimension du signal vectoriel, et $r_{one-shot}(\underline{y})$ est le débit total de la structure monorésolution. Ainsi, cette structure est légèrement sous-optimale relativement à une approche monorésolution à cause de retour de bruit lié à la structure de prédiction en boucle fermée.

Finalement, cette structure à deux étages est généralisée au cas de M étages. Dans ce cas, les débits associés aux différentes résolutions sont contrôlés par des quantificateurs de pas Δ_{Q_k}. Nous proposons la règle suivante pour calculer les Δ_{Q_k} afin que les débits r_k de chaque résolution approchent des débits cibles R_k prédéterminés:

$$\Delta_{Q_k} \approx \left[2^{r_{scal,L(z)}(\underline{y}) - \sum_{i=1}^{k} R_i} \right]_1 , \quad k = 1, ..., M. \tag{10.27}$$

Des exemples numériques évaluant la qualité de cette méthode sont finalement présentés. Ces travaux sont présentés dans [144].

10.4 Conclusions

Cette thèse propose diverses techniques de codage avec et sans pertes pour les signaux vectoriels. Ces techniques sont présentées comme les divers aspects d'un cadre théorique général basé sur la notion de causalité. Les performances des divers systèmes de codage proposés sont analysées aux moyens d'outils statistiques et, pour la plupart, sous l'hypothèse Gaussienne. Les résultats théoriques ont été confrontés à des simulations numériques et sont résumés dans cette partie.

La première partie de cette thèse a présenté des techniques de codage avec perte pour les signaux vectoriels.

Dans le cadre du codage par transformée (CT) tout d'abord, nous nous sommes intéressés au codage de signaux vectoriels par une transformation décorrélatrice causale de type DPCM (Differential Pulse Code Modulation, technique utilisée pour les signaux scalaires, supprimant les redondances par prédiction linéaire). Nous avons montré que la transformation causale optimale correspond à une factorisation triangulaire LDU (Lower-Diagonal-Upper) de la matrice d'autocorrélation du vecteur de signal à coder. Cette approche a été ensuite comparée à sa contrepartie unitaire, la transformation de Karhunen-Loève (KLT), bien connue parce qu'étant optimale pour les sources Gaussiennes, elle sert traditionnellement de référence. Plusieurs aspects sont abordés dans cette comparaison, comme le gain de codage apporté par la transformation (qui correspond au facteur par lequel la distorsion est réduite, pour un même débit, grâce à la transformation), les effets intervenants lorsque le schéma de codage est implémenté en boucle fermée (c'est à dire lorsque la transformation utilise des données précédemment quantifiées, ce qui introduit dans le schéma de codage un retour de bruit), ou la complexité algorithmique. Nous avons proposé une analyse des perturbations liées au retour de bruit, qui montre que quand celui-ci devient négligeable, les performances sont identiques à celles obtenues dans le cas unitaire, bien que la complexité de la LDU soit notablement moindre. Ainsi, cette transformation apparaît comme un modèle optimal alternatif à la traditionnelle transformation de Karhunen-Loève.

Dans la plupart des cas pratiques cependant, les données réelles ne sont pas stationnaires, ce qui pose un problème d'adaptation pour des transformations dépendant du signal, telles que la KLT ou la LDU. Nous avons donc cherché à étudier les performances de schémas de codage dont les paramètres sont adaptés "en ligne" (sur la base de données quantifiées uniquement), ce qui évite de transmettre un surcroît de débit associé à ces paramètres. Dans ce contexte, nous avons analysé les effets de perturbation liés au bruit de quantification et au bruit d'estimation qui se posent par rapport au cas idéal où la matrice de covariance est connue parfaitement. Sous certaines hypothèses simplificatrices empruntées au CT classique, cette analyse a permis d'évaluer quantitativement, en fonction d'un débit moyen imposé et du nombre de données précédemment décodées, l'écart entre la performance réelle des deux systèmes et leur performance idéale, où les statistiques des signaux à compresser sont connues.

Poursuivant l'analyse de systèmes de CT "en ligne", nous nous sommes tournés vers l'analyse de systèmes concrets utilisant des quantificateurs uniformes suivis de codeurs entropiques, pour lesquels le mécanisme d'allocation de bits est simple, et proche de l'optimalité. Les résultats de cette partie ont montré que des systèmes adaptatifs (à pas de quantification fixes ou adaptatifs) peuvent fournir des performances similaires à des systèmes conçus avec une connaissance *a priori* de la source, bien qu'aucune information concernant les transformations ou le pas de quantification utilisé ne soit transmise au décodeur. Ces analyses traitent du cas causal comme du cas unitaire.

Dans la fin de cette première partie, l'approche matricielle causale de type LDU a été généralisée au cas où les coefficients de la matrice de transformation triangulaire sont des filtres prédicteurs (prédiction MIMO -Multi Input Multi Output). Cette généralisation a débouché sur la prédiction MIMO dite "généralisée", pour indiquer que la prédiction MIMO classique et la prédiction MIMO triangulaire constituent deux cas particuliers, parmi une infinité, d'une même approche totalement décorrélatrice, et "causale" dans un sens plus large. Pour la prédiction triangulaire, la causalité correspond à l'ordre dans lequel les signaux sont décorrélés. Comme pour la LDU, nous avons analysé le gain de codage sous une hypothèse de résolution infinie d'abord; les effets de retour de bruit de quantification ont ensuite été pris en compte. Nous avons montré que pour la prédiction MIMO triangulaire, décorréler les signaux par ordre de variance décroissante est optimal. Une application de ces résultats a été proposée dans le cadre du codage de la parole large bande ([0-7kHz]).

La deuxième partie de cette thèse a développé des techniques de codage sans perte basées sur les approches causales considérées précédemment.

Une première étape a consisté à comparer les performances de la LDU à celles de la KLT dans le cas où elles sont implémentées de façon à être sans perte (transformations d'entiers à entiers). Le gain correspond alors à la réduction de débit opérée par la transformation (par rapport à un codage entropique scalaire direct des coefficients quantifiés), tout en garantissant une représentation exacte de la source. Nous avons montré d'abord que le gain maximal qui peut être apporté par de telles transformations correspond à la moyenne des informations mutuelles partagées par les différentes variables qui composent le processus vectoriel. Nous avons ensuite analysé les gains apportés par la KLT et la LDU dans ce cadre, et avons décrits les effets dûs aux non linéarités (contrainte "entiers à entiers") en terme de débit supplémentaire par rapport au cas idéal. Le bruit d'estimation pour un schéma adaptatif a aussi été traité. L'approche causale, grâce à sa nature triangulaire, s'avère présenter dans ce cadre des performances légèrement supérieures à l'approche unitaire.

Nous avons ensuite étudié des schémas de codage sans perte qui permettent de délivrer, dans un premier temps, une version basse résolution du signal d'intérêt, et de transmettre le signal complémentaire par la suite. Ce genre de schémas est utile pour des applications de navigation rapide sur internet, ou de transmission à bande passante variable par exemple. La transformation causale a été comparée dans ce cadre aux transformations orthogonales. Nous avons considéré un schéma à deux niveaux de résolution simple (utilisé par exemple dans le contexte du codage audio sans perte), dans lequel chaque vecteur est d'abord transformé, quantifié, puis transmis comme version basse résolution du signal. Un signal d'erreur est ensuite généré par soustraction au signal original, et transmis comme complément. L'extension de ce schéma à plusieurs niveaux de résolution a été obtenue en introduisant des quantificateurs de type APCM dans le schéma sans perte. On a montré que les transformations orthogonales classiques sont sous-optimales pour de telles approches multirésolution par rapport à leur alternative causale. La transformation causale présente d'autres avantages par rapport à des transformations telles que la KLT ou la DCT, comme la possibilité de passer instantanément d'un schéma de codage sans pertes monorésolution à un schéma multirésolution, de pouvoir choisir des niveaux de résolution différents pour chacun des canaux et, notamment, de pouvoir coder sans pertes un ou plusieurs canaux particuliers uniquement. Finalement, des schémas de codage sans pertes multirésolutions ont été proposés, qui se basent sur la prédiction MIMO considérée dans la première partie. Nous avons montré que l'approche multirésolution est légèrement sous-optimale en terme de débit total par rapport à une approche de compression globale à cause du retour de bruit dans les boucles de type ADPCM. On a aussi proposé une méthode pour que les débits générés par chacune des résolutions correspondent à des débits cibles prédéterminés.

Comme cela transparaît dans le résumé ci-dessus, de nombreuses techniques de codage ont été considérées dans ce travail [8], notamment le CT, le codage en sous-bandes, les transformations d'entiers à entiers ou le codage multirésolution. Le choix d'un large champ d'investigations est néanmoins à double tranchant. D'un côté, un large panorama était nécessaire pour décrire l'étendue, la diversité et l'intérêt théorique des approches causales. D'un autre côté, chacun des thèmes traités a dégagé des questions intéressantes, mérite certainement des approfondissements. Pour des applications pratiques, les systèmes considérés peuvent être améliorés et complexifiés, même si dans ce cas une modélisation théorique peut devenir difficile. Dans le cas du CT "en ligne" par exemple, une étude approfondie de ces systèmes devrait inclure le choix d'un quantificateur adapté à l'application considérée, ainsi qu'aux signaux à coder; le problème de l'adaptation des fenêtres temporelles pose aussi d'intéressant problèmes pratiques comme théoriques. En ce qui concerne les prédicteurs MIMO et notamment le prédicteur triangulaire, les performances de systèmes pratiques, basés sur des filtres RIF, dépendra fortement d'un choix adéquat du nombre de coefficients dédiés à la décorrélation intersignaux. Le degré d'anticausalité dédié à ces filtres devrait être optimisé relativement à la longueur des trames, ou relativement à un délai de reconstruction dans l'optique d'un codeur échantillon par échantillon. Par ailleurs, une question importante dans le codage de source est celle de critères subjectifs: si l'évaluation de l'erreur quadratique moyenne est un critère simple, et qui permet de mener facilement des analyses théoriques, elle renseigne souvent très mal sur la qualité effectivement perçue d'un codeur audio ou d'images... Enfin, d'un point de vue théorique, il semble que deux axes de recherches se dessinent naturellement à la suite de ce travail. Premièrement, il serait intéressant de rechercher s'il y a d'autres (ou la classe de toutes) les transformations qui, comme la KLT et la LDU, sont optimales pour des sources Gaussiennes. Deuxièmement, la modélisation de performances des systèmes considérés dans cette thèse gagnerait en intérêt si, au moyen de mixtures de Gaussiennes, elle pouvait décrire des sources de densités de probabilité arbitraires.

[8] En dehors du codage de source, la prédiction MIMO généralisée s'est avérée utile en detection multi-utilisateurs [53].

Bibliography

[1] J. Y. Huang and P. Schultheiss, “Block quantization of correlated Gaussian random variables,” *IEEE Trans. on Comm.*, vol. 11, pp. 289–296, Sept. 1963.

[2] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, continued October 1948.

[3] T. M. Cover and J. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.

[4] J. P. Perez, “L’entropie de Boltzmann et l’entropie de Shannon, même concept ?,” *Bulletin de l’Union des Physiciens*, vol. 92, pp. 145–155, June 1998.

[5] E. Jaynes, “Information theory and statistical mechanics,” *Physical review*, vol. 166, no. 4, p. 620, 1957.

[6] A. Reeves. French Patent 852 183, October 3 1938.

[7] J. Pierce, “The early years of information theory,” *IEEE Trans. on Inf. Th.*, vol. IT19(1), pp. 3–8, January 1973.

[8] H. Black, “Pulse code modulation,” *Bell Lab. Record.*, vol. 25, pp. 265–269, July 1947.

[9] H. Black and E. Odson, “PCM equipment,” *Elec. Eng.*, vol. 66, pp. 1123–1125, November 1947.

[10] V. Nourrit, *Etude des fonctions de routage spatial et fréquentiel en espace libre. Application à la conception de fonctions optiques pour les télécommunications*. PhD thesis, Université de Bretagne Occidentale, June 2002.

[11] N. Moreau, *Techniques de compression des signaux*. Collection technique et scientifique des communications, Masson, 1995.

[12] G. Gonon, *Proposition d’un schéma d’analyse/synthèse adaptatif dans le plan temps-fréquence basé sur des critères entropiques. Application au codage audio par transformée*. PhD thesis, Université du Maine, June 2002.

[13] S. Vembu, S. Verdu, and Y. Steinberg, “The source-channel separation theorem revisited,” *IEEE trans. on Inf. Th.*, vol. 40, pp. 44–54, January 1995.

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

[15] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[16] R. Gray and D. Neuhoff, “Quantization,” *IEEE trans. on Inf. Th.*, vol. 44, October 1998.

[17] T. Berger and J. Gibson, “Lossy source coding,” *IEEE Trans. on Inf. Th.*, vol. 44, pp. 2693–2723, Oct. 1998.

[18] D. A. Huffman, “A method for the construction of minimum redundancy codes,” in *IRE*, (Los Angeles, CA), pp. 1098–1101, Sept. 1952.

[19] R. G. Gallager, “Variations on a theme by Huffman,” *IEEE trans. on Inf. Th.*, vol. 22, pp. 668–674, Nov. 1978.

[20] R. M. Capocelli and A. DeSantis, “Variations on a theme by Gallager,” in *Image and Text Compression*, (Boston), pp. 181–213, J.A. Storer Ed.,Kluwer, 1992.

[21] M. Purat, T. Liebchen, and P. Noll, “Lossless transform coding of audio signals,” in *102nd AES Convention, Munich*, 1997.

[22] A. Bruekers, A. Oomen, and R. van der Vleuten, “Lossless coding for DVD audio,” in *101st AES Convention*, (Los Angelese, CA), Nov. 1996. Preprint 4358.

[23] T. Robinson, “Shorten: simple lossless in near-lossless waveform compression,” Tech. Rep. 156, Cambridge University Eng. Dep., UK, 1994.

[24] M. Hans and R. Schafer, “Lossless Compression of Digital Audio,” *IEEE Sig. Proc. Mag.*, vol. 18, Jul. 2001.

[25] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” *IRE National Convention Records, Part 4*, pp. 142–163, 1959.

[26] B. M. Oliver, J. Pierce, and C. Shannon, “The philosophy of PCM,” in *Proc. IRE*, vol. 36, pp. 1324–1331, Nov. 1948.

[27] W. Bennett, “Spectra of quantized signals,” *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, 1948.

[28] W. Panter and W. Dite, “Quantization distortion in pulse-count modulation with nonuniform spacing of levels,” in *Proc. IRE*, vol. 39, pp. 44–48, Jan. 1951.

[29] V. Koshelev, “Quantization with minimal entropy,” *Probl. Pered. Inform.*, vol. 14, pp. 151–156, 1993.

[30] T. Goblick and J. Holsinger, "Analog source digitization: A comparison of theory and practice," *IEEE Trans. Inf. Th.*, pp. 323–326, Apr. 1967.

[31] R. Wood, "On optimum quantization," *IEEE Transactions on Information Theory*, vol. IT-15, pp. 248–252, Mar. 1969.

[32] S. Lloyd, "Least squares quantization in PCM." Bell Lab. Tech. Note, Sept 1957.

[33] S. Lloyd *: IEEE Trans. on Inf. Th., Special Issue on Quantization*, pp. 129–137, Mar. 1982.

[34] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inf. Th.*, vol. IT-6, pp. 7–12, Mar. 1960.

[35] H. Gish and J. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inf. Th.*, vol. IT-14, pp. 676–683, Sept. 1968.

[36] T. Berger, "Optimum quantizers and permutation codes," *IEEE Trans. on Inf. Th.*, pp. 759–765, Nov 1972.

[37] R. Zelinski and P. Noll, "Investigations on quantization of memoryless Gaussian sources," tech. rep., Heinrich Hertz Institut, Berlin, 1972.

[38] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Math. Acad. Sci. Hungar.*, vol. 10, pp. 193–215, 1959.

[39] H. Kramer and M. Mathews, "A linear coding for transmitting a set of correlated signals," *IRE*, vol. 23, pp. 41–46, Sept. 1956.

[40] R. A. Horn and C. Johnson, *Matrix analysis*. Cambridge University Press, 1985. Th. 7.8.1.

[41] V. Goyal, "Transform coding with integer-to-integer transforms," *IEEE trans. on Inf. Theory*, vol. 46, March 2000.

[42] K. Karhunen, "Über lineare Methoden in der Wahrscheinlichkeitsrechnung," *Ann. Acad. Sci. Fenn., Ser. A1,: Math.-Phys.*, vol. 37, pp. 3–79, 1947.

[43] M. Loève, *Processus stochastiques et mouvements Browniens*, ch. Fonctions aléatoires de second ordre. P. Levy, Ed. Paris, France:Gauthier-Villars, 1948.

[44] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychology*, vol. 24, pp. 417–441, 498–520, 1933.

[45] V. Goyal, "High-rate transform coding: How high is high, and does it matter?," in *Proc. IEEE Int. Symp. Inform Theory*, (Sorrento, Italy), p. 207, June 2000.

[46] M. Effros, H. Feng, and K. Zeger, "Suboptimality of Karhunen-Loève transform for transform coding." Submitted to: IEEE Trans. Inf. Theory.

[47] V. Goyal, “Theoretical foundations of transform coding,” *IEEE Signal Processing Magazine*, vol. 18, pp. 9–21, Sept. 2001.

[48] V. Goyal, “Single and multiple description transform coding with bases and frames.” Philadelphia, PA: SIAM, 2001.

[49] M. Effros, H. Feng, and K. Zeger, “Suboptimality of Karhunen-Loève transform for transform coding,” in *DCC*, 2003.

[50] M. Effros, “Rate-distortion bounds for fixed- and variable-rate multiresolution sources codes.” Submitted to IEEE Trans. on Inf. Theory on March 26, 1998.

[51] D. J. Sakrison, “Worst sources and robust codes for difference distortion measures,” *IEEE Trans. Inf. Theory*, vol. 21, pp. 301–309, May 1975.

[52] K. P. anf K. Zeger, “Robust quantization of memoryless sources using dispersive FIR filters,” *IEEE Trans. Comm.*, vol. 40, Nov. 1992.

[53] A. Medles and D. Slock, “Linear convolutive space-time precoding for spatial multiplexing MIMO systems,” in *39th Annual Allerton Conference on Communication, Control, and Computing*, (Monticello, Illinois, USA,), Oct. 2001.

[54] V. Goyal, J. Zhuang, and M. Vetterli, “Transform coding with backward adaptive updates,” *IEEE Transactions on Information Theory*, vol. 46, July 2000.

[55] D. Mary and D. T. M. Slock, “Codage DPCM vectoriel et application au codage de la parole en bande Élargie,” in *CORESA 2000*, (Poitiers, France), October 2000.

[56] D. Mary and D. T. M. Slock, “Vectorial DPCM coding and application to wideband coding of speech,” in *ICASSP 2001*, (Salt Lake City), May 2001.

[57] D. Mary and D. T. M. Slock, “Comparison between unitary and causal approaches to transform coding of vectorial signals,” in *GRETSI01 - Corrected Version, available at http://www.eurecom.fr/˜mary/publications.html*, (Toulouse, France), September 2001.

[58] D. Mary and D. T. M. Slock, “Comparison between Unitary and Causal Approaches to Backward Adaptive Transform Coding of Vectorial Signals,” in *Proc.ICASSP 2002*, (Orlando, USA), May 2002.

[59] D. Mary and D. T. M. Slock, “Backward adaptive transform coding of vectorial signals : A comparison between unitary and causal approaches,” in *EUSIPCO-2002*, (Toulouse, France), September 2002.

[60] D. Mary and D. T. M. Slock, “Rate-distortion analysis of practical backward adaptive transform coding schemes,” in *ICASSP03*, (Hong-Kong).

[61] D. Mary and D. Slock, “Causal transform coding, generalized MIMO prediction and application to vectorial DPCM coding of multichannel audio,” in *WASPAA01*, (New York, USA), October 2001.

[62] S. Phoong and Y. Lin, “PLT versus KLT,” in *IEEE Int. Symp. Circ. Syst.*, May 1999.

[63] S.-M. Phoong and Y.-P. Lin, “Prediction-based lower triangular transform,” *IEEE trans. on Sig. Proc.*, vol. 48, July 2000.

[64] F. Lahouti and A. Khandani, “Sequential vector decorrelation technique,” Tech. Rep. 2001-4, Univ. of Waterloo, March 2001.

[65] A. Gersho and R. Gray, *Vector quantization and signal compression*. Kluwer Academic, 1992.

[66] N. Levinson, “The wiener RMS (root mean squared) error criterion in filter design and prediction,” *J. Math. Phys.*, vol. 25, pp. 261–278, 1947.

[67] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ: Prentice Hall, 3rd. ed., 1996.

[68] D. Arnstein, “Quantization errors in predictive coders,” *IEEE trans. on Communications*, pp. 423–429, April 1975.

[69] R. W. Stroh, *Optimum and Adaptive Differential Pulse Code Modulation*. PhD thesis, Polytechnic Inst., Brooklyn, N.Y., 1970.

[70] A. Sripad and D. Snyder, “A necessary and sufficient condition for quantization errors to be uniform and white,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 442–448, Oct. 1977.

[71] B. Widrow, I. Kollar, and M.-C. Liu, “Statistical theory of quantization,” *IEEE Transactions on Instrumentation and Measurement*, vol. 45, pp. 353–361, Apr. 1996.

[72] L. Cheded and P. Payne, “The exact impact of amplitude quantization on multi-dimensional, high-order moments estimation,” *Signal Processing*, vol. 39, 1994.

[73] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[74] D. L. Neuhoff, R. M. Gray, and L. D. Davisson, “Fixed rate universal block source coding with a fidelity criterion,” *IEEE Trans. Inf. Theory*, vol. IT-21, pp. 511–523, Sept. 1975.

[75] P. A. Chou, M. Effros, and R. Gray, “A vector quantization approach to universal noiseless coding and quantization,” *IEEE Trans. on Inf. Theory*, vol. 42, pp. 1109–1138, July 1996.

[76] M. Effros and P. A. Chou, “Weighted universal transform coding: Universal image compression with the Karhunen-Loève transform,” in *Proc. Int. Conf. Image Processing*, vol. II, pp. 61–64, Oct. 1995.

[77] B. Ottersten and M. Viberg, “Sensor Array Signal Processing,” tech. rep., Royal Institute of Tech., Chalmers Technical Inst., Sweden, Juanuary 1994.

[78] H. Lütkepohl, *Introduction to Multiple Time Series Analysis.* New-York : Springer Verlag, 1993.

[79] P. P. Vaidyanathan, *Multirate Systems and Filter Banks.* Englewood Cliffs, NJ: Prentice Hall, 1993.

[80] R. de Queiroz and H. Malvar, "On the asymptotic performances of hierarchical transforms," *IEEE Trans. Sig. Proc.*, vol. 40, pp. 2620–2622, Oct. 1992.

[81] S. Rao and W. Pearlman, "Analysis of linear prediction, coding, and spectral estimation from subbands," *IEEE Trans. on Inf. Th.*, vol. 42, pp. 1160–1178, 1996.

[82] C. Giurcaneanu, I. Tabus, and J. Astola, "Linear prediction from subbands for lossless audio compression," in *NORSIG'98, IEEE Nordic Signal Processing Symposium*, pp. 225–228, 1998.

[83] T. Fischer, "On the rate-distortion efficiency of subband coding," *IEEE Transactions on Information Theory*, vol. 38, March 1992.

[84] P. Vaidyanathan, "Quadrature mirror filter banks, M-band extensions and perfect reconstruction techniques," *IEEE ASSP Mag.*, pp. 4–20, Jul. 1987.

[85] B. Maison and L. Vanderdorpe, "About the asymptotic performance of multiple-input/multiple-output linear prediction of subband signals," *IEEE Signal Processing Letters*, vol. 5, December 1998.

[86] L. Vanderdorpe and B. Maison, "Multiple input/multiple output linear prediction of subbands signals," *IEEE Trans. on Circ. and Syst.*, vol. 46, pp. 1230–1233, Sept. 1999.

[87] P. W. Wong, "Rate distortion efficiency of subband coding with crossband prediction," *IEEE Transactions on Information Theory*, Januar 1997.

[88] S. Tate, "Band ordering in lossless compression of multispectral images," in *DCC*, pp. 311–320, 1994.

[89] A. Rao and S. Bhargava, "Multispectral data compression using bidirectional interband prediction," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 34, pp. 385–397, March 1996.

[90] V. Cuperman and A. Gersho, "Adaptive differential vector coding of speech," in *IEEE GLOBECOM*, pp. 1092–1096, Dec. 1982.

[91] C. W. Rutledge, "Vector predictive coding of color images," in *IEEE GLOBECOM*, pp. 1158–1164, Dec. 1986.

[92] J. Fowler, M. Carbonara, and S. Ahalt, "Image coding using differential vector quantization," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 3, pp. 350–367, Oct. 1993.

[93] D. Yang, H. Ai, C. Kyriakakis, and C.-C. Kuo, "An inter-channel redundancy removal approach for high-quality multichannel audio compression," in *AES 109th Convention*, (Los Angeles, USA), September 2000.

[94] J. Pearl, “On coding and filtering stationary signals by discrete Fourier transform,” *IEEE Transactions on Information Theory*, pp. 229–232, March 1973.

[95] D. Yang, *High Fidelity Multichannel Audio Compression*. PhD thesis, University of Southern California, May 2002.

[96] D. Bauer and D. Seitzer, “Statistical properties of high quality stereo signals in the time domain,” in *ICASSP*, pp. 2045–2048, May 1989.

[97] S.-S. Kuo and J. Johnston, “A study of why cross channel prediction is not applicable to perceptual audio coding,” *IEEE Sig. Proc. Letters*, vol. 8, pp. 245–247, Sept. 2001.

[98] V. Jain and R. Crochiere, “Quadrature mirror filter design in the time domain,” *IEEE Trans. on Acoust. Speech and Sig. Proc.*, vol. ASSP-32, pp. 353–361, April 1984.

[99] M. Smith and T. Barnwell, “Exact reconstruction techniques for tree-structured subband coders,” *IEEE Trans. on Acoust. Speech and Sig. Proc.*, vol. ASSP-34, pp. 434–441, June 1986.

[100] P. Jax and P. Vary, “An upper bound on the quality of artificial bandwidth extension of narrowband speech signals,” in *ICASSP*, pp. 237–240, 2002.

[101] D. Mary and D. Slock, “Evaluating the efficiency of joint optimisation for formant and pitch predictors,” tech. rep., Institut Eurécom, Sept. 2000. Available at http://www.eurecom.fr/ mary/publications.html.

[102] D. Mary and D. Slock, “Rapport final : Projet RNRT Cobasca - travaux d’Eurécom,” tech. rep., Institut Eurécom, Ap. 2001.

[103] P. Burt and E. Adelson, “The Laplacian pyramid as compact image code,” *IEEE Trans. Commun.*, vol. COM-31, pp. 532–540, Apr. 1983.

[104] M. Bellanger, G. Bonnerot, and M. Coudreuse, “Digital filtering by polyphase networks: application to sample rate alteration and filterbanks,” *IEEE Trans. on Acoust. Speech and Sig. Proc.*, vol. ASSP-24, pp. 109–114, April 1976.

[105] M. Viergever and P. Roos, “Hierarchical interpolation, an efficient method for reversible compression of images,” *IEEE Engineering in Medicine and Biology*, pp. 48–54, March 1993.

[106] A. Netravali and B. Haskell, *Digital Pictures: Representation and Compression*. Plenum Press, New York, 1988.

[107] M. Hans and R. Schafer, “Lossless compression of digital audio,” *IEEE Signal Processing Magazine*, vol. 18, July 2001.

[108] T.Moriya, “Report of AHG on Issues in Lossless Audio Coding,” tech. rep., $mpeg - audiolossless@research.att.com$, December 2001. ISO/IEC JTC1/SC29/WG11 M7678.

[109] R. Whittle, “Lossless audio compression.” URL : http://www.firstpr.com.au/audiocomp/lossless/.

[110] T. Moriya, “Reflector for AHG on Issues in Lossless Audio Coding,” tech. rep., $mpeg - audiolossless@research.att.com$, December 2001.

[111] V. Goyal, “Multiple description coding: compression meets the network,” *IEEE Signal Processing Magazine*, vol. 18, pp. 74–93, Sept. 2001.

[112] V. Koshelev, “Hierarchical coding of discrete sources,” *Probl. Pered. Inform*, vol. 16, pp. 31–49, 1980.

[113] W. Equitz and T. Cover, “Successive refinement of information,” *IEEE Trans. on Information Theory*, vol. 37, pp. 269 –275, Mar. 1991.

[114] B. Rimoldi, “Successive refinement of information: characterization of the achievable rates,” *IEEE Trans. on Inform.Theory*, vol. 40, pp. 253 –259, Jan. 1994.

[115] D. Pan, “A tutorial on MPEG/audio compression,” *IEEE Multimedia Journal*, pp. 60–74, 1995. Summer Issue.

[116] B. Bessette, R. Lefebvre, R. S. M. Jelinek, J. Vainio, J. R.-P. andH. Mikkola, and K. Järvinen, “Techniques for high-quality ACELP coding of wideband speech,” in *Eurospeech*, 2001.

[117] G. T. 26.190, “Adaptive multi-rate wideband speech transcoding,” 2001. 3GPP Technical Specification.

[118] Y. Huang, H. Dreizen, and N. Galatsanos, “Prioritized DCT for compression and progressive images,” *IEEE Trans. Im. Proc.*, vol. 1, Oct. 1992.

[119] N. Boulgouris, “Optimal progressive lossless image coding using reduced pyramids with variable decimation ratios,” *IEEE Trans. on Im. Proc.*, vol. 9, Dec. 2000.

[120] G. Gagnon, “Multiresolution video coding for HDTV,” in *Canadian Conference on Electrical and Computer Engineering*, vol. 1, pp. 19–22, 1993.

[121] S. Fukuma, M. Iwahashi, and N. Kambayashi, “Adaptive multi-channel prediction for lossless scalable coding,” in *ISCAS*, vol. 4, pp. 467 –470, 1999.

[122] A. Benazza-Benyahia, J.-C. Pesquet, and M. Hamdi, “Lossless coding for progressive archival of multispectral images,” in *ICASSP*, vol. 3, pp. 1817 –1820, 2001.

[123] M. Iwahashi, S. F. ans S. Chokchaitam, and N. Kambayashi, “Lossless/lossy progressive coding based on reversible wavelet and lossless multi-channel prediction,” in *ICIP*, vol. 1, pp. 430 –434, 1999.

[124] N. Jayant, “Variable rate ADPCM based on explicit noise coding,” *Bell System Tech. J.*, pp. 655–677, Mar. 1983.

[125] S. Zhang and G. Lockhart, “An efficient embedded ADPCM coder,” in *Telecommunications, IEE*, no. 404, March 1995.

[126] C. R. G.727, “5-,4-,3-, 2 bit/sample embedded ADPCM,” Jul. 1990.

[127] M. H. Sherif, D. Bowker, G. Bertocci, B. Orford, and G. Mariano, “Overview and performance of CCITT/ANSI embedded ADPCM algorithms,” *IEEE Trans. on Com.*, vol. 41, no. 2, pp. 391–399, 1993.

[128] T. Nomura, M. Iwadare, M. Serizawa, and K. Ozawz, “A bitrate and bandwidth scalable CELP coder,” in *ICASSP*, pp. 341–344, 1998.

[129] R. de Queiroz and J. Yabu-Uti, “Hierarchical image coding with pyramid DPCM,” in *SBT/IEEE International Telecommunications Symposium Record.*, pp. 558–562, Sept. 1990.

[130] F. Amano, K. Iseda, K. Okazaki, and S. Unagami, “An 8 kbps TC-MQ (time domain compression ADPCM-MQ) speech codec,” in *ICASSP*, pp. 259–262, April 1988.

[131] J. Shapiro, “Embedded image coding using zerotrees of wavelets coefficients,” *IEEE Trans. Sig. Proc.*, pp. 3445–3462, Dec. 1993.

[132] A. Said and W. Pearlman, “A new, fast and efficient image codec based on set partioning in hierarchical trees,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 243–250, June 1996.

[133] Wu and Memon, “Context-based, adaptative, lossless image coding,” *IEEE Trans. on Comm.*, vol. 45, pp. 437–444, Apr. 1997.

[134] T. Ramabadran and K. Chen, “The use of context information in the reversible compression of medical images,” *IEEE Trans. on Medical Imaging*, vol. 11, pp. 185–195, June 1992.

[135] I. Avcibas, N. Memon, B. Sankur, and K. Sayood, “A progressive lossless/near-lossless image compression algorithm,” *IEEE Sig. Proc. Letters*, vol. 9, pp. 312–314, Oct. 2002.

[136] T. Moriya, A. Jin, and T. Mori, “A design of lossy and lossless scalable audio coding,” in *ICASSP*, pp. 889–892, 2000.

[137] T. Moriya, A. Jin, T. Mori, K.Ikeda, and T.Kaneko, “Lossless scalable audio coder and quality enhancement,” in *ICASSP*, vol. 2, pp. 1829–1832, 2002.

[138] R. Öktem, . N. Gerek, E. Cetin, L. Öktem, and K. Egiazarian, "Adaptive filter banks for lossless image compression," in *ICASSP01*, 2001.

[139] N. Boulgouris, D. Tzovaras, and M. G. Strintzis, "Lossless image compression based on optimal prediction, adaptive lifting, and conditional arithmetic coding," *IEEE Trans. on Im. Proc.*, vol. 10, Jan. 2001.

[140] P. Roos, M. Viergever, M. VanDijke, and J. Peters, "Reversible intraframe compression of medical images," *IEEE Trans. on Medic. Imag.*, vol. 7, pp. 328–336, Dec. 1988.

[141] G. Kuduvalli and R. Rangayyan, "Performance analysis of reversible image compression techniques for high-resolution digital teleradiology," *IEEE Trans. on Medic. Imag.*, vol. 11, Sept. 1992.

[142] D. Mary and D. T. M. Slock, "A performance analysis of integer-to-integer transforms for lossless coding of vectorial signals," in *Proc. SMMSP'02*, (Toulouse, France), September 2002.

[143] D. Mary and D. T. M. Slock, "On the suboptimality of orthogonal transforms for single- or multi-stage lossless transform coding," in *DCC*, 2003.

[144] D. Mary and D. T. M. Slock, "Multistage integer-to-integer multichannel prediction for scalable lossless coding," in *Asilomar Conf. on Signals, Systems and Computers*, (Pacific Grove, CA), Nov. 2002.

[145] H. Blume and A. Fand, "Reversible and irreversible image data compression using the S-Transform and Lempel-Ziv coding," in *SPIE*, vol. 1091, pp. 2–18, 1989.

[146] A. Zandi, J. Allen, E. Schwartz, and M. Boliek, "CREW: Compression with Reversible Embedded Wavelets," in *DCC*, pp. 212–221, March 1995.

[147] A. Said and W. Pearlman, "An image multiresolution representation for lossless and lossy compression," *IEEE Trans. on Im. Proc.*, vol. 5, pp. 1303–1310, Sept. 1996.

[148] M. Adams, F. Kossentini, and R. K. Ward, "Generalized S transform," *IEEE Trans. on Sig. Proc.*, vol. 50, Nov. 2002.

[149] W. Sweldens, "The lifting scheme: a custom design construction of biorthogonal wavelets," *J. Appl. Compt. Harmon. Anal.*, vol. 3, no. 2, pp. 186–200, 1996.

[150] S. Dewitte and J. Cornelis, "Lossless integer wavelet transform," *IEEE Sig. Proc. Letters*, vol. 4, pp. 158–160, June 1997.

[151] A. Calderbank, I. Daubechies, W. Sweldens, and B.-L. Yeo, "Wavelet transforms that map integers to integers," *Applied and Computational Harmonics Analysis*, vol. 5, no. 3, pp. 332–339, 1998.

[152] I. Daubechies and W. Sweldens, “Factoring wavelets transforms into lifting steps,” *J. Fourier Anal. Appl.*, vol. 4, no. 3, pp. 247–269, 1998.

[153] Q. Shi, “Biorthogonal wavelets theory and techniques for image coding,” in *SPIE*, pp. 24–32, Oct. 1998.

[154] S. Oraintara, Y.-J. Chen, and T. Nguyen, “Integer fast Fourier transform,” *IEEE Trans. Sig. Proc.*, vol. 50, pp. 607–618, March 2002.

[155] L. J and T. Tran, “Fast multiplierless approximations of the DCT with the lifting scheme,” *IEEE Trans. on Sig. Proc.*, vol. 49, Dec. 2001.

[156] W. Chen, C. Harrison, and S. Fralick, “A fast computational algorithm for the discrete cosine transform,” *IEEE Trans. on Com.*, vol. COM-25, no. 9, pp. 1004–1011, 1977.

[157] C. Loeffler, A. Lightenberg, and G. Moschytz, “Practical fast 1D DCT algorithm with 11 multiplications,” in *ICASSP*, vol. 2, pp. 988–991, Feb. 1989.

[158] J. Hong, *Discrete Fourier, Hartley and cosine transforms in signal processing*. PhD thesis, Columbia University, Department of Electrical Engineering, 1993.

[159] A. Patel and M. Tonkelowitz, “Lossless sound compression usind the discrete wavelet transform.” Available at http://www.eecs.harvard.edu/ vernal/academic/publications.shtml, Jan. 2002.

[160] P. Hao and Q. Shi, “Matrix factorizations for reversible integer mapping,” *IEEE Trans. on Sig. Proc.*, vol. 49, Oct. 2001.

[161] C. Giurcaneanu and I. Tabus, “Low-complexity transform coding with integer-to-integer transforms,” in *ICASSP*, pp. 2601 –2604, 2001.

[162] I. Csiszàr, “Generalized entropy and quantization problems,” in *6th Prague Conf.*, pp. 159–174, 1973.

[163] J. Stuart, P. Craven, M. Gerzon, M. Law, and R. Wilson, “MLP lossless compression,” in *AES 9th Convention*, (Tokyo, Japan), 1998.

[164] Y. Wang, L. Yaroslavsky, M. Vilermo, and M. Vaananen, “A multichannel audio coding algorithm for inter-channel redundancy removal,” in *110th AES Convention, Amsterdam*, May 2001.

[165] S. Oraintara, Y.-J. Chen, and T. Nguyen, “Integer fast Fourier transform,” *IEEE trans. on Signal Processing*, vol. 50, March 2002.

[166] R. M. Gray, “Toeplitz and Circulant Matrices: A Review,” tech. rep., Stanford University, California, 1971, revised August 2002. Available at http://ee-www.stanford.edu/ gray/toeplitz.html.

[167] T. Liebchen, “Lossless audio coding using adaptive multichannel prediction,” in *113th AES Convention, Los Angeles*, October 2002.