OPTIMAL ORTHONORMAL SUBBAND CODING AND LATTICE QUANTIZATION WITH VECTOR DITHERING

Thesis by

Ahmet Kıraç

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1999

(Submitted July 31, 1998)

Acknowledgements

I would like to thank my advisor Prof. P. P. Vaidyanathan for all the guidance and support he has provided. I have greatly benefited from his academic excellence and his style of thinking. He has always managed to bring more clarity and soundness to any idea that we have discussed together. Thank you P. P., thanks for everything you have taught me.

I would like to thank Prof. Robert J. McEliece, Dr. Marvin K. Simon, Prof. Jehoshua Bruck, and Dr. Jalil Fadavi for serving on my defense committee. Dr. Simon has been very kind to show high level of interest in both of my defense and candidacy presentations. I enjoyed the penetrating questions of Prof. McEliece in both of these presentations. The encouraging comments of Prof. Bruck are appreciated. Dr. Fadavi is from Lucent Technologies, my next destination. I would like to thank him for his kind interest and warm support.

I thankfully recall the good times with my colleagues Dr. See-May Phoong, Dr. Yuan-Pei Lin, Dr. Igor Djokovic, Dr. Jamal Tuqan, Murat Meşe, and Sony Akkarakaran. Jamal has been my buddy and a very good company for all the years. I thank him for his limitless interest and curiosity on many technical issues that enabled us to have very productive and inspiring discussions. Murat has brought the lab charm and energy, and Sony has brought calmness and serenity. Outside the lab, I have been fortunate to have as friends, the wonderful people, Ayhan İrfanoğlu and Zehra Çataltepe. I cannot describe my appreciation and thankfulness towards them with words. I would like to also acknowledge the friendship and the supports of Dr. Doruk Engin, Murat Somer, Dr. Slim Alouini, and Vildana Gatskich. Thanks also to our system administrator Robert Freeman, the secretaries Lavonne Martin, Tanya Hefner, Linda Dozsa, and Lilian Porter. I extend my best wishes to our librarian Paula Samazan.

My family is my best blessing for which I am thankful to God. I would like to acknowledge the love and support of my parents, my five brothers, and my only sister.

Abstract

In the digital era that we live in, efficient coding of signals is an unquestionable need. This thesis is about one of the most useful and popular technique of digital coding: subband coding. Subband coding and its cousin wavelet-based coding are now the preferred methods for not only speech, but also audio, image, and video signals. Subband coding involves a linear part which is a filter bank, and a nonlinear part which is usually a uniform scalar quantization of each of the subbands. Subband coders are classified according to the type of filter bank used for its transform. This thesis is mainly about orthonormal subband coding. The ability of an orthonormal filter bank to decompose the signal into components that have a diverse set of signal energies is an indicator of its efficiency for subband coding. Such a diversity in the set of the subband energies is fully utilized by a process called bit allocation. The traditional results on the optimality of a filter bank for given input statistics assume that the quantizers operate at high bit rates.

This thesis presents optimality results under more general quantizer models without assuming high bit rates. This is accomplished by revealing the relationship between the problems of optimal orthonormal subband coding and principal component representation of signals. The latter is done using what is called a principal component filter bank (PCFB). A PCFB is one that compacts most of the energy of a signal into smaller subsets of subbands. To date, there has not been significant theoretical developments in the field of optimal nonuniform subband coding, although the successful techniques of wavelet-based coding are among the state of the art in practice. Such techniques utilize a form of a nonuniform filter bank with a certain structure which makes it efficient for its implementation. In this thesis, we provide optimality results for the nonuniform orthonormal subband coding as well. As in the uniform case, the principal component representation of signals continues to play the key role. We introduce nonuniform PCFB's and link them to the optimal subband coding problem. A

PCFB, in particular, contains a filter that compacts most of the signal energy into one single channel: energy compaction filter. The thesis goes into details of designing such filters optimally. In particular, we propose an analytical method in the two-channel case and a very efficient window method in the arbitrary $M-$channel case. Multistage design of compaction filters has also been worked out.

Finally we extend the analysis of uniform scalar quantization to multiple dimensions. We provide an exact statistical relationship between a lattice quantizer noise and its input vector. We then extend the idea of dithering to the vector case. Dithering is a means of statistically rendering the quantization noise independent of the input. We address the optimal choice of a lattice for a given dimension and also optimal pre- and post-filtering of a dithered lattice quantizer.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Digital Signal Processing (DSP) has undoubtedly penetrated into the heart of modern technology. Digital domain is not only the preferred domain for the storage and retrieval of a signal of any type, but it is also the preferred domain for its processing and transmission. The transition between the analog domain and the digital domain is done using analog-to-digital (A/D) and digital-to-analog (D/A) converters. An analog signal $x(t)$ is both continuous-time and continuous-amplitude signal, and it is converted into a digital signal $x(n)$ which is both discrete-time and discrete-amplitude. Time domain conversion is through the process of **sampling** while the amplitude domain conversion is done via **quantization**. The signals in this thesis are of digital type; however, on many occasions they can be considered as discrete-time but continuous-amplitude.

Within DSP lies the world of multirate signal processing. Sampling and quantization of continuous signals lead to the notion of **resolution** in time and amplitude. The higher the sampling frequency and the higher the number of quantization levels, the better the resolution of the digital approximations of the original signals. The cost is of course the increase in the volume of data that we have to deal with. Multirate signal processing strives to keep that cost at reasonable levels by applying smart signal processing algorithms that involve alterations of sampling rates and the number of quantization levels. A central topic in multirate signal processing is **subband coding**

which involves filter banks and quantization of subband signals.

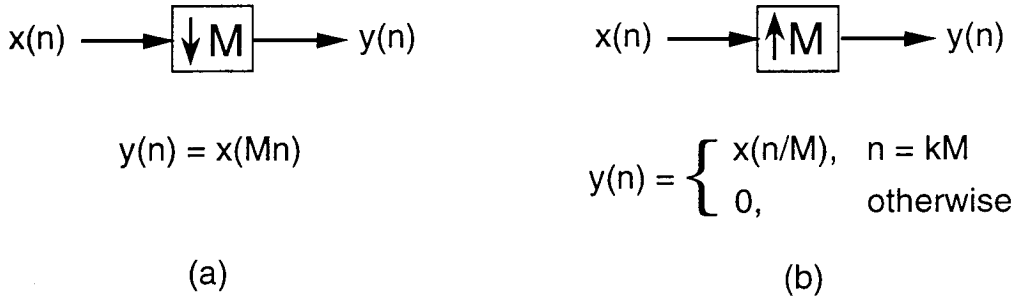


Fig. 1.1: (a) Downsampling (b) Upsampling.

The term *multirate* originated in applications that involve sampling rate alterations of signals. Essential to these applications are the operations of downsampling and upsampling as shown in Fig. 1.1. In words, downsampling keeps every Mth sample while upsampling inserts $M-1$ zeros between consecutive samples.

A fundamental concept in any technical field is the concept of transformation. Depending on the application, the original signal is transformed into a more convenient domain, processed in that domain, and then transformed back into the original domain. A transformation can also be viewed as decomposition into a basis and the inverse transformation viewed as reconstruction using the transform coefficients. In traditional digital signal processing, an example of a transformation is a linear time-invariant (LTI) system (a filter) represented by its transfer function $H(z)$ as in Fig. 1.2. The filter produces an output sample for each input sample. A second example is the matrix

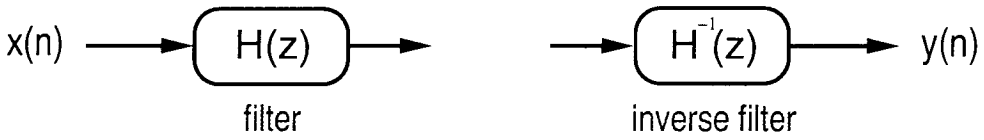


Fig. 1.2: Traditional signal processing: an LTI system and its inverse.

transform of a vector signal as illustrated in Fig. 1.3. The transform produces an output vector for each input vector.

In the world of multirate signal processing, the transformations are done by filter banks as in Fig. 1.4. In the figure, the original signal $x(n)$ is passed through M different

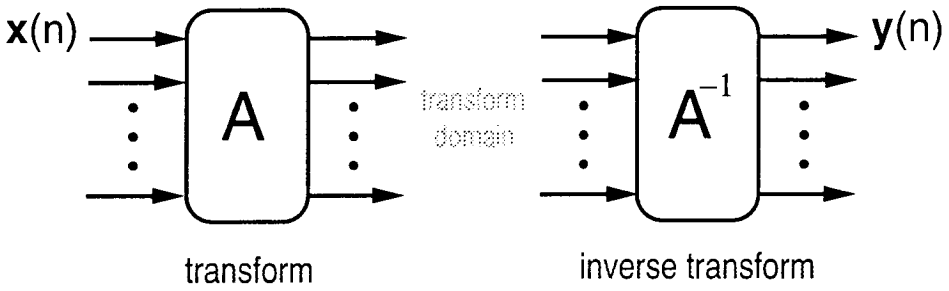


Fig. 1.3: Transformation of a vector signal.

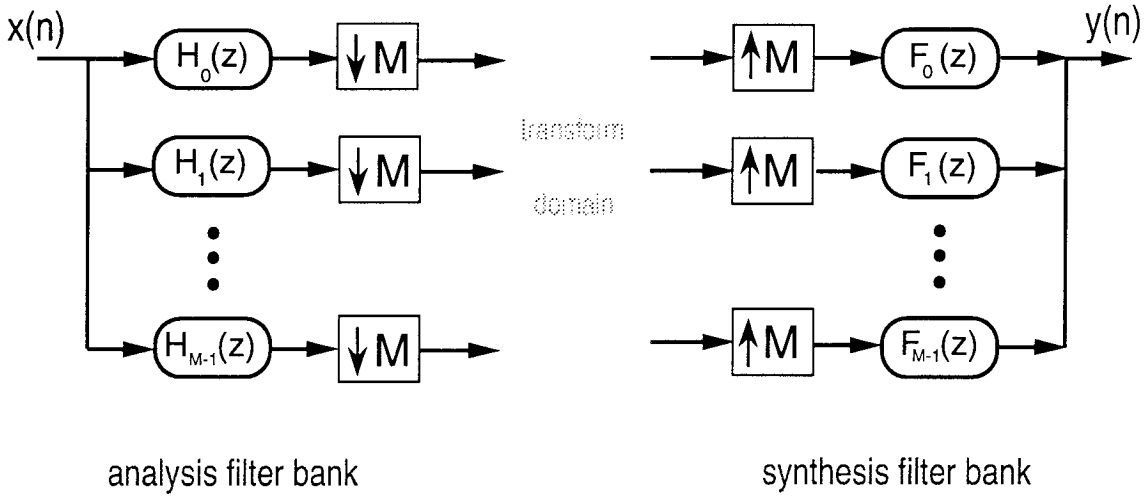


Fig. 1.4: Multirate signal processing: analysis and synthesis filter bank.

filters. We have an M-fold increase in the number of signals, so we decimate each output signal by M. Similarly, after processing of these M components, we upsample each of them by M and pass them through filters before combining into the output signal $y(n)$ of the same sampling rate as the original signal. The set of filters that are used to decompose the original signal into M channels is called the **analysis** filter bank, and the set of filters that are used to recombine the processed channels is called the **synthesis** filter bank.

A typical filter bank has a frequency response as shown in Fig. 1.5. Thus each of the subband signals corresponds to a different portion of the frequency spectrum of the input signal. Subband signals are time sequences on their own. Hence in a sense, filter banks produce joint time and frequency representations of signals.

In most applications, it is desirable that the transformation be invertible. In Fig. 1.2

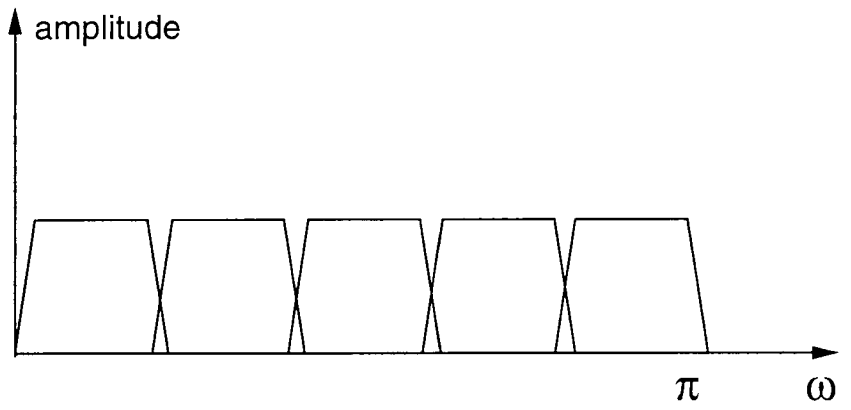


Fig. 1.5: Frequency response of a typical filter bank.

the inverse transformation of $H(z)$ is simply the inverse LTI filter $H^{-1}(z)$ and it is an easy task to design $H(z)$ such that its inverse is well-behaved, that is, it is stable. Similarly, in Fig. 1.3 the inverse transformation is simply $\mathbf{A}^{-1}$ as long as the matrix $\mathbf{A}$ is nonsingular. In Fig. 1.4, it is not a trivial matter to design a set of filters for the synthesis filter bank that form the inverse of the analysis filter bank. Whenever this happens, we say that the system has the perfect reconstruction (PR) property. In Fig. 1.2, either the filter $H(z)$ or its inverse $H^{-1}(z)$ has to be recursive (IIR). A beautiful result in filter bank theory is that it is possible to have a PR system with analysis and synthesis filter banks that both have nonrecursive (FIR) filters [Vai93].

A fundamental tool in filter bank theory is the concept of polyphase representation. The idea is to represent the filter bank system by a multi-input-multi-output (MIMO) transfer function that operates on a vector signal. The vector signal is obtained by blocking of the original scalar signal as shown in Fig. 1.6.

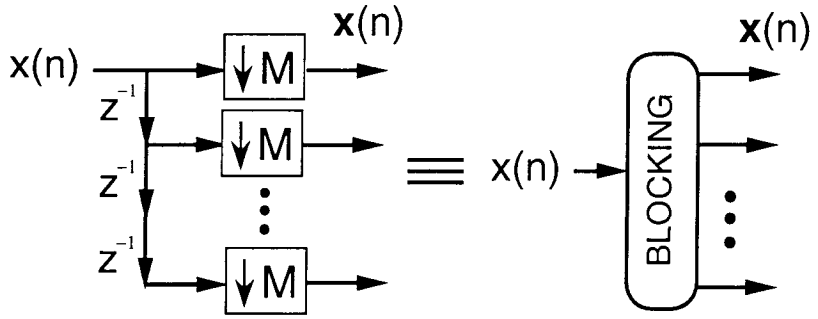


Fig. 1.6: Blocking of the scalar input.

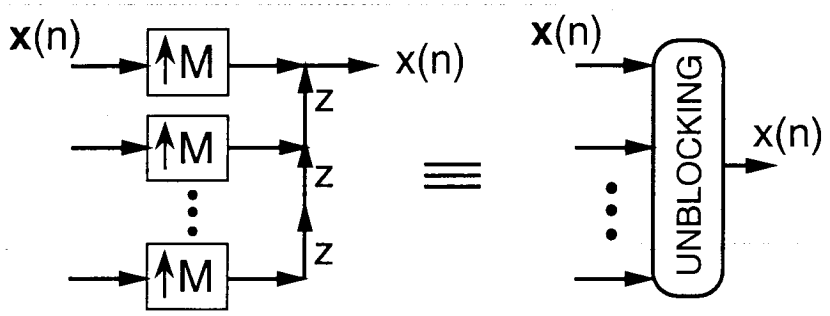


Fig. 1.7: Unblocking of the vector input.

Once this fundamental connection is done, many of the questions in filter bank theory can be brought into the realm of MIMO systems. It has been shown that the analysis filter bank of Fig. 1.4 is equivalent to blocking followed by a MIMO system as shown in Fig. 1.8. In a sense, sequential processing of the input by M filters is

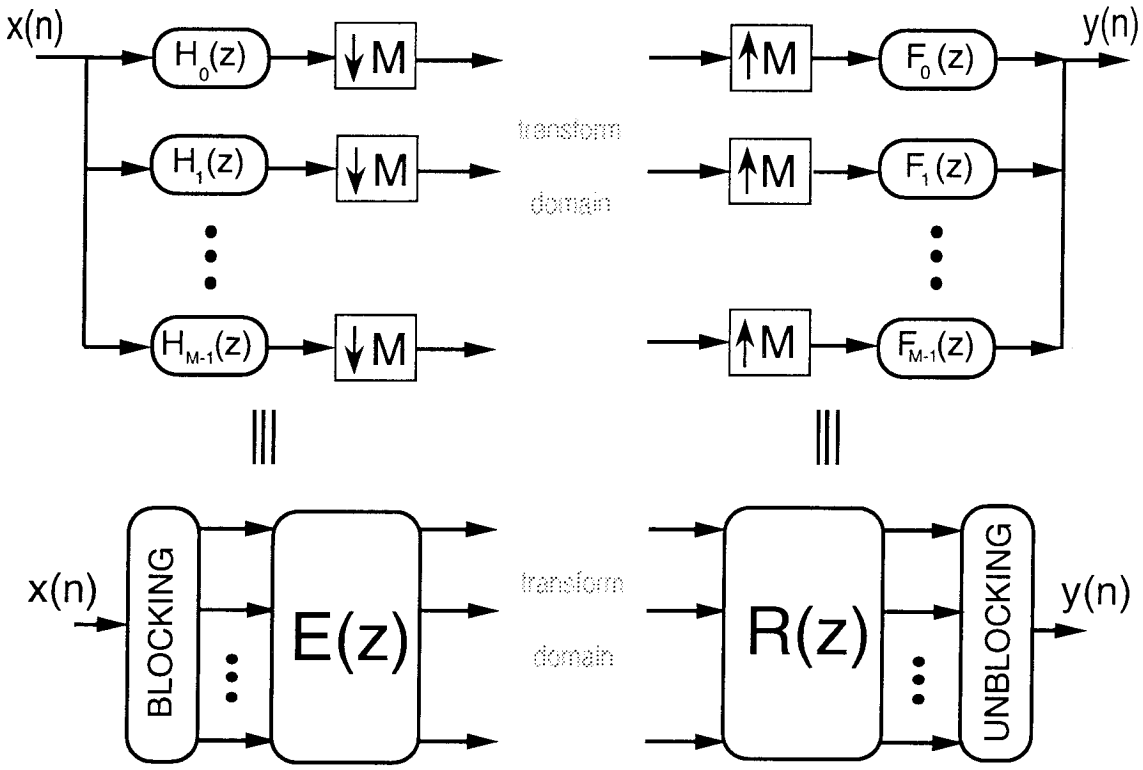


Fig. 1.8: Polyphase representation of the analysis filter bank.

replaced by the parallel processing of the blocks of the input by a MIMO system. This is more efficient in terms of implementation because all the processing is done at lower sampling rates. Similarly, we have the equivalence shown in Fig. 1.8 for the synthesis

filter bank, where the unblocking is the reverse of the blocking operation as shown in Fig. 1.7. Comparing Fig. 1.2 and Fig. 1.4, we see that a filter bank can be viewed as a natural extension of a classical filter. Comparison of Fig. 1.3 with Fig. 1.8 puts in evidence the fact that a filter bank is also a natural extension of a matrix transform. While for some applications it may prove more efficient to view the filter banks as extensions of transforms, i.e., MIMO systems, there are many occasions where they can be best thought of as multiple filters operating at different portions of a single frequency spectrum, e.g., as in spectral analyzers.

1.1 Classification of Filter Banks

Consider Fig. 1.4 again. Each of the M subband signals is decimated by the same number M resulting in an average sampling rate that is the same as the input sampling rate. Such a filter bank is called a **uniform filter bank**. Equivalently one can think of using different decimation ratios for different channels while keeping the average sampling rate the same. If n_i denotes the decimation ratio for ith channel, then this can be accomplished by having $\sum_{i=0}^{M-1} 1/n_i = 1$. Such a filter bank is called a **nonuniform filter bank**. Fig. 1.9 shows frequency responses of a typical nonuniform filter bank. The theory and design of nonuniform filter banks are relatively less developed than the

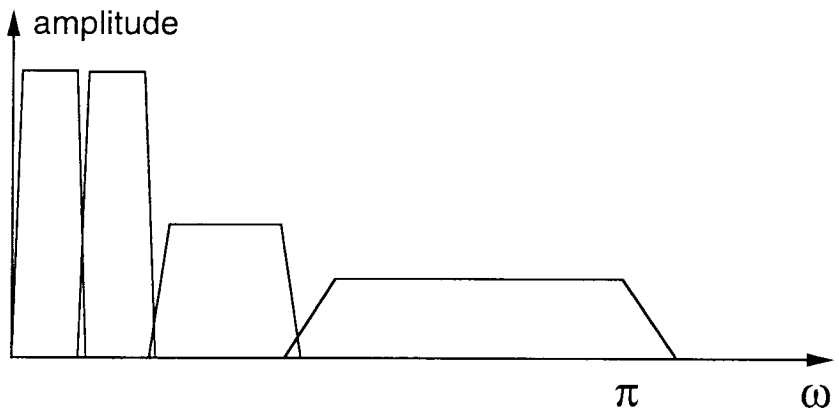


Fig. 1.9: Frequency response of a typical nonuniform filter bank.

uniform counterparts. For the uniform case, the classification of filter banks can be

done easily in terms of the equivalent MIMO systems $\mathbf{E}(z)$ and $\mathbf{R}(z)$ in their polyphase representation.

1. **Biorthogonal filter bank.** A filter bank with perfect reconstruction (PR) property. In terms of the polyphase representation, the MIMO systems should satisfy:

$$\mathbf{R}(z)\mathbf{E}(z) = \mathbf{I} \tag{1.1}$$

In terms of the filter transfer functions, the equivalent condition is

$$H_i(z)F_j(z)\Big|_{\downarrow M} = \delta(i-j) \tag{1.2}$$

The reader should refer to Sec. 1.4 on notations to make sense of (1.2).

2. **Orthonormal filter bank.** A biorthogonal filter bank with further condition :

$$\mathbf{R}(z) = \tilde{\mathbf{E}}(z) \tag{1.3}$$

In terms of filter transfer functions, we have equivalently

$$F_i(z) = \tilde{H}_i(z) \tag{1.4}$$

The notation $\tilde{}$ for matrix and scalar transfer functions is explained in Sec. 1.4. In the time domain, (1.4) is equivalent to $f_i(n) = h_i^*(-n)$. Hence in the orthonormal case, the synthesis filters can be determined from the analysis filters by time reversal and complex conjugation.

3. **Transform.** We call the filter bank simply a transform if $\mathbf{E}(z) = \mathbf{E}_0$, a constant matrix. In this case, the operation of the filter bank is simply a block by block matrix transformation of the original signal. In terms of filters, a transform is equivalent to a filter bank with filter orders $< M$.

4. **FIR filter bank.** A filter bank with FIR filters. In terms of the polyphase representation, an analysis filter bank is FIR if

$$\mathbf{E}(z) = \mathbf{E}_0 + \mathbf{E}_1 z^{-1} + \ldots + \mathbf{E}_K z^{-K} \tag{1.5}$$

In contrast to a transform, an FIR filter bank can be thought of as a lapped transform. The current output vector is the summation of matrix transformations of the current and K previous input blocks. Here K is called the **order** of the MIMO system $\mathbf{E}(z)$. There is another significant notion, similar to, but different from the order. It is the **degree** of the MIMO system $\mathbf{E}(z)$ which represents the minimum number of delay elements required to implement it.

5. **Ideal filter bank.** A filter bank with no order constraints on the filters. The polyphase transfer function $\mathbf{E}(e^{j\omega})$ can be arbitrarily set for each ω. The filters might be of infinite length and noncausal. The current output vector is the summation of matrix transformations of all input blocks.

1.2 Filter Banks For Subband Coding

Subband coding (SBC) is perhaps the most important application of filter banks. Even before the theory of filter banks was developed, SBC was innovated as an efficient technique of encoding speech signals. Subband coding and its cousin wavelet-based coding are now the preferred methods for not only speech, but also audio, image, and video signals. We show in Fig. 1.10 a subband coding scheme using a uniform filter

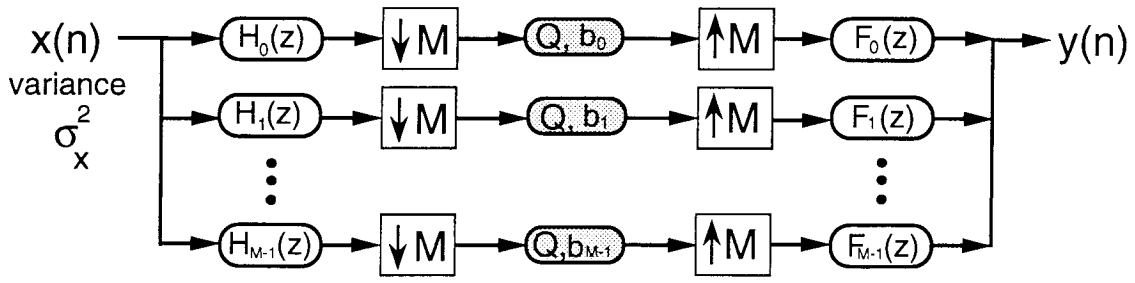


Fig. 1.10: Subband coding scheme with a uniform filter bank.

bank. The subband signals are quantized using different numbers of bits. The original signal has a variance σ_x^2 and the subband signals have variances $\sigma_{x_i}^2$, $i = 0, 1, \ldots, M-1$. Essential to a SBC scheme is the process of **bit allocation**. The success of subband coding is highly dependent on the way the bit allocation is performed. An intuitive way to accomplish this task is to assign more bits to the channels with higher energies as illustrated in Fig. 1.11. In the extreme case, if after applying a PR filter bank, some of

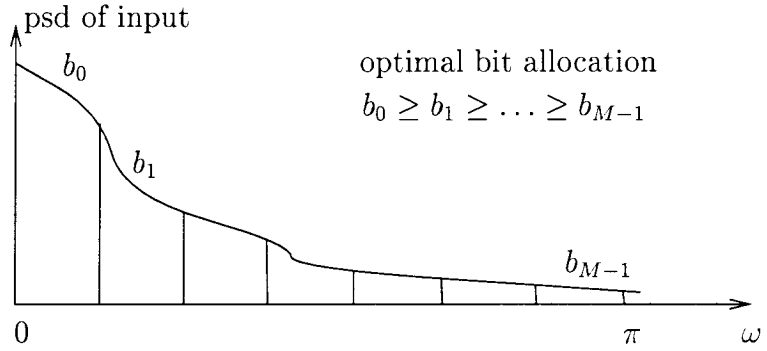


Fig. 1.11: Bit allocation according to the subband variances.

the channels are identically zero, then those channels need not be assigned any bits at all. The advantage of using a filter bank is then to be able to treat different portions of the signal (in the frequency domain) differently, as opposed to the scalar quantization where the same number of bits are assigned to every single sample. Of course, one can adapt the number of bits in the time domain according to how much energy exists in various periods of time. By using a filter bank, we can do such an adaptation for each of the channels. In a sense, with subband coding, it is possible to take advantage of the variations of the signal both in time and frequency domains. This is the underlying fundamental reason for the success of recently introduced wavelet-based image coding technique called the **zero-tree coding** [Sha93, SP96]. In this technique, the bits are allocated in an adaptive fashion jointly in the space domain (equivalent to time domain) and in the scale domain (analogous to frequency domain). The final destination for audio visual signals is the human perception. Hence it makes sense to understand and use the properties of the human ear and the human visual system (HVS). It turns out that subband coding schemes are ideally suitable for taking the perceptual properties into account.

Even though we choose the filter bank to have the perfect reconstruction property, the subband quantizers introduce noise and lead to a distortion between the input $x(n)$ and the reconstructed output $y(n)$. The type of distortion that we consider throughout the thesis is the mean-squared error, that is,

$$\mathcal{E} = E[|x(n) - y(n)|^2] \tag{1.6}$$

By the uniformity of the filter bank we have

$$b = \sum_{i=0}^{M-1} b_i/M \tag{1.7}$$

Hence an **optimal subband coder** is the one that minimizes the distortion (1.6) for a fixed average bit rate (1.7). Optimal subband coding involves both designing the filter bank and doing the allocation of bits in an optimal fashion. In the special orthonormal case, we have

$$\sigma_x^2 = \sum_{i=0}^{M-1} \sigma_{x_i}^2/M \tag{1.8}$$

In the further special case that the quantizers operate at high bit rates, the quantizer noise variances are $\sigma_{q_i}^2 = c2^{-2b_i}\sigma_{x_i}^2$ and the minimum distortion for any filter bank is of the form

$$\mathcal{E} = c2^{-2b}\left(\prod_{i=0}^{M-1} \sigma_{x_i}^2\right)^{1/M} \tag{1.9}$$

If the assumption that the quantizers operate at high bit rate does not hold, then the above expression for the distortion is not valid. In this thesis we will present results pertaining to this arbitrary bit rate case. In order to show what has been known in the area of optimal subband coding and what has been done in this thesis, we have the summary chart shown in Fig. 1.12. In the chart we presented the solved cases with dark lines. A brief history that goes along with the chart is given next.

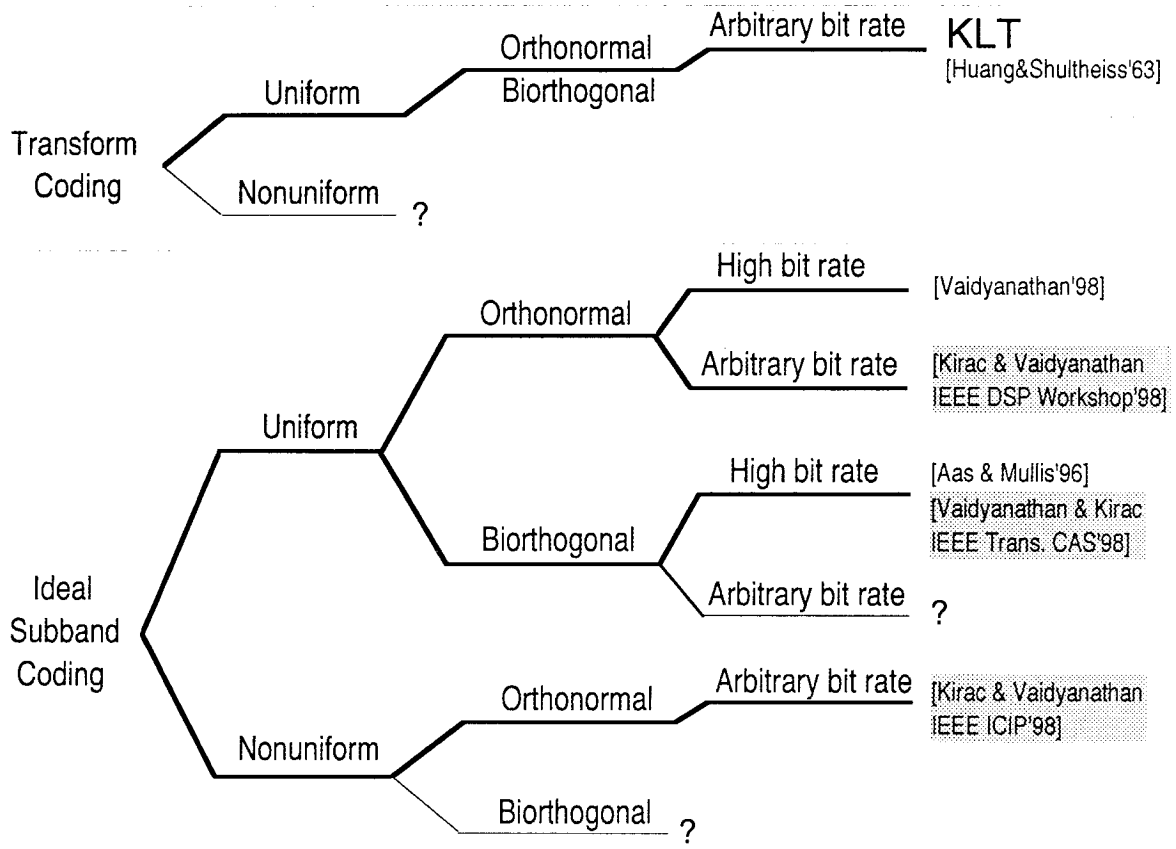


Fig. 1.12: Summary of the known and the new results in optimal subband coding.

1.2.1 A Brief History of Optimal Subband Coding

Theoretical results on the optimization of filter banks for subband coding have been first obtained for the transform coding case. The first pioneering work is due Huang and Shultheiss in 1963 [HS63]. They have shown in [HS63] that Karhunen-Loeve transform (KLT) is the optimal transform coder under mild assumptions on the quantization noise sources. In particular, they have shown the optimality of KLT even when the high bit rate assumptions do not hold. Surprisingly, the extension of transform coding to the nonuniform case has not even been considered! Of course, one has to define first what a nonuniform transform is. Looking at the uniform transform case where the equivalent filters have lengths less than M, one plausible definition of nonuniform transform is that the equivalent filter for the ith channel has length less than n_i, where n_i is the decimation ratio for that channel.

Going back to the uniform case, optimal orthonormal filter banks of unconstrained filters (ideal filter banks) have recently been constructed by Vaidyanathan [Vai98] under the assumption that the quantizers operate at high bit rates. The theory of optimal ideal biorthogonal filter banks under the high bit rate assumptions have also been considered recently by Aas and Mullis [AM96].

In the nonuniform case, surprisingly there has not been significant theoretical developments. In contrast, many of the state of the art coding algorithms used in practice utilize some form of nonuniform filter banks. A good example is the zero-tree coding, a wavelet-based image coding technique as proposed by Shapiro in 1993 [Sha93] and later improved by Said and Pearlman in 1996 [SP96]. The transform in this technique is a dyadic tree structured filter bank which is a special form of a nonuniform filter bank.

Going back again to the uniform case, we have just mentioned the theoretical developments for the transform coding case and the ideal subband coding case. These two can be considered as two extreme cases and the intermediate case would be the SBC schemes with FIR filter banks. There had not been significant theoretical developments in this case either. On the practical side, Malvar has invented algorithms to design efficient filter banks which he called lapped orthonormal transforms (LOT) [Mal92]. He later extended these designs to the biorthogonal case (LBT) [Mal97].

1.3 Thesis Overview

1.3.1 Contributions in Optimal Subband Coding

One of the main contributions of the thesis is the discovery of the connection between optimal orthonormal subband coding problem and the principal component representation of signals. Consider Fig. 1.13 where we show a filter bank with P of its subband channels directly connected to the synthesis part, while the remaining channels are disconnected. In the synthesis part, the disconnected subband signals are replaced with zeros. In this scheme the source of distortion between the input and the output is not

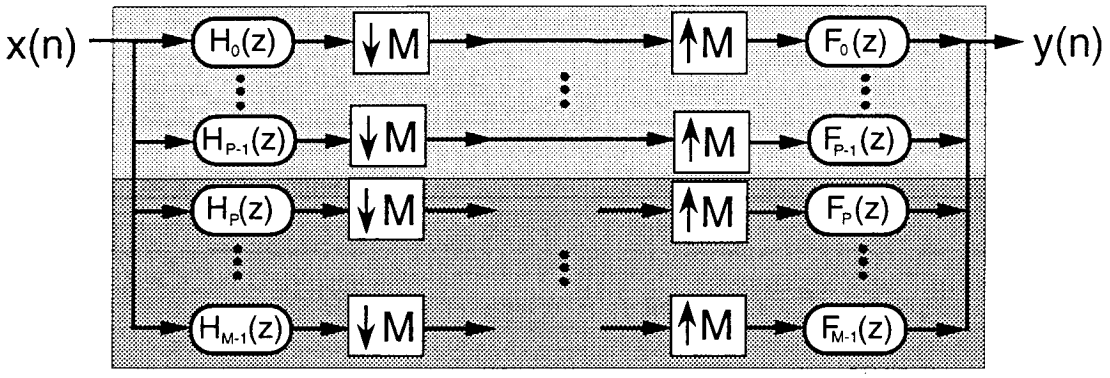


Fig. 1.13: Pertaining to the introduction of principal component filter banks.

due to the quantizers but the fact that some of the subband signals are dropped. If we keep only one channel, then we have only one analysis and synthesis filter pair to consider. If we keep two channels, we have two pairs of filters, and so on. A **principal component filter bank (PCFB)** is the one that minimizes the distortion between the input and the output for any number of retained channels (See Chapter 2 for the precise definition). For a given input statistics, principal component filter banks with unconstrained filters are constructed by Tsatsanis and Giannakis [TG95]. It turns out that this solution coincides with the solution of optimal orthonormal filter banks as constructed in [Vai98] with the high bit rate assumptions on quantizers. In Chapter 2 of this thesis, we show that this is not just a mere coincidence and that the two problems are fundamentally connected to each other. In particular we show that PCFB's are optimal orthonormal filter banks for subband coding and that the optimality is valid at arbitrary bit rates as well.

As we pointed out before, there has been no significant theoretical developments in the nonuniform subband coding case. Chapter 2 of this thesis contains the first results for this case. The chapter introduces the extension of principal component filter banks to the nonuniform case. First, a fixed set of decimation ratios are assumed for the construction of optimal nonuniform orthonormal filter banks, and then the question of optimum set of decimation ratios for a fixed number of channels is addressed.

In Chapter 2, we also address the case of FIR orthonormal subband coding. The fundamental connection between PCFB's and optimal orthonormal filter banks con-

tinues to exist; however, there is the question of the existence of PCFB's in the FIR case. With FIR constraints, there are optimal sets of filter pairs that minimize the reconstruction error in Fig. 1.13 for any number of retained channels. The problem is, there is no single filter bank with FIR constraints that minimizes the reconstruction error for all retained number of channels.

Although in the special transform coding case there is no advantage of using a biorthogonal transform [Vai98], the situation is different for the general subband coding case. A construction of optimal uniform biorthogonal filter banks is advanced in [AM96]. However, their proof of optimality seems not to be complete as we address in [VK98a]. We show in [VK98a] that with a further assumption of uniqueness of the solution, such a construction is indeed optimal. The formulation of the problem has a certain symmetry in terms of the analysis and the synthesis filter banks, and it is this symmetry that enables us to give a complete proof of optimality. The optimal solution has two stages: the input signal is first passed through an optimal uniform orthonormal filter bank (or a PCFB) and then each of the decorrelated channels are individually filtered with the so-called half-whitening filters (see [VK98a] for details). We also provide in [VK98a] some theoretical bounds on the coding gain of a biorthogonal subband coder.

1.3.2 Optimal FIR Compaction Filters

Although PCFB's in FIR case do not always exist, it is still a good idea to design a pair of filters ($P = 1$ in Fig. 1.13) to minimize the reconstruction error. In doing this we somehow pack most of the energy into one single channel. One can then think of completing this pair of filters to a filter bank. If we constrain the filter bank to be orthonormal and keep its degree fixed, then completion to a filter bank turns out to be a straightforward procedure that involves canonical factorization of the polyphase vector corresponding to the compaction filter [MM98]. In the orthonormal case it suffices to design the analysis filters only, as the synthesis filters are determined from the analysis ones (see (1.4)). The major question is then how to design the first FIR

filter in an optimal fashion. Such filters are called **optimal FIR compaction filters** and are treated in detail in Chapter 3 of this thesis. The chapter gives a good overview of the existing relevant work followed by new results and design strategies. Here are the major contributions of this chapter:

1. **Analytical method for the two-channel case.** In the two-channel case, the design of an FIR compaction filter has practical as well as theoretical significance. The theoretical aspect lies in the fact that the second filter of a two-channel orthonormal filter bank is determined from the first one by simple flipping and sign changes in time domain [Vai93]. This fact has the exceptional implication that the problems of optimal compaction filter and optimal orthonormal subband coding are the same, even with the order constraints! Another implication is that FIR PCFB's do exist in the two-channel case as opposed to the arbitrary M-channel case. Hence it would be a significant result if we could somehow be able to determine optimal FIR compaction filters in an analytical fashion. For a restricted class of input statistics, we derive such an analytical technique in Sec. 3.3.

2. **Window Method.** In $M-$channel case, there is no analytical technique that we are aware of. An optimal $M-$channel FIR compaction filter can be completed to an $M-$channel orthonormal filter bank, though it would not be the optimal one. This is because of the fact that, in general, $M-$channel FIR PCFB's do not exist. Hence it is a loss of generality to design the first filter to be a compaction filter. Since we cannot obtain an optimal FIR filter bank from an optimal compaction filter, it would be good idea to trade off the efficiency of the design of a compaction filter with its optimality. We have come up with such an efficient method, and it is presented in detail in Sec. 3.4. The design technique which we called the window method involves FFT and simple comparison.

3. **Multistage designs.** If the number of channels M is a composite number, then it is possible to design compaction filters in multiple stages, each of which involving the design of reduced size compaction filters. The effective order of

the resulting compaction filters are much higher than the orders of individual filters. One form of such a multistage design uses linear programming technique, while another form uses any of the available design methods for compaction filters, linear programming being only one of them. In Sec. 3.5 we also propose a method to improve the linear programming technique itself for designing compaction filters.

1.3.3 Lattice Quantization and Vector Dithering

As we have mentioned, quantization forms an essential part of digital signal processing, in particular, multirate signal processing. The subband coding schemes utilize the simplest types of quantizers, namely, uniform scalar quantizers. There are of course much more efficient ways of quantizing, examples of which include Lloyd-Max quantizers, entropy constrained vector quantizers, etc. Although the quantizers in the subbands are primitive, effectively a subband coding scheme is like a sophisticated vector quantizer. If the filter bank in a subband coding scheme is designed well, then as long as the subband signals are quantized independently, the benefit of using sophisticated quantizers for each channel is only marginal. In the recently developed coding algorithms like zero-tree coding, the quantization is effectively a uniform scalar one with a dead-zone (meaning the bin for the quantization level of zero is wider than the other uniform bins). However, the allocation of bits is somewhat adaptive and the adaptation is done jointly across the subbands and within the subbands.

Since uniform quantizers turn out to be very important components of practical algorithms, we wanted to analyze their statistical behavior given the statistics of the input. In the scalar case, it did not take us too much time to discover the wonderful works of Schuchman [Sch64] and Sripad and Snyder [SS77]. While the former had the analysis of so called dithered quantizers, the latter was about the analysis of scalar uniform quantization itself. Dithering was an interesting technique that was invented and used in image coding by Roberts [Rob62]. The idea is to add a pseudo random sequence v to the quantizer input x and, if possible, subtract it in the reconstruction

side. By such a simple technique, it is possible to render the quantization error $e = x - \hat{x}$ independent of the input signal x. The practical implication of this is a superior perceptual quality of quantized signals. In a sense, the resolution of quantized signals is increased perceptually without using extra quantization levels.

If the pseudo random sequence is not subtracted in the reconstruction side for some reason, then we have what is called a nonsubtractive dithering scheme. Although the probability density function (pdf) of a pseudo random sequence v does not play a role in the subtractive case, it does so in the nonsubtractive case, as the quantization error depends on it. The investigation of the optimal pdf that minimizes the error was our first research project. The optimal pdf turned out to be a triangular distribution. The width of the support of the pdf is the same as twice the quantization step size. We then realized that nonsubtractive dithering with such an optimal dither pdf was already in use in the audio industry.

In Chapter 4, we extend the analysis of uniform scalar quantization and dithering to multiple dimensions: lattice quantization and vector dithering. In the chapter, we demonstrate the perceptual advantages of vector dithering, and propose techniques to design suitable pseudo random vector sequences. It turns out that lattice quantization error is dependent on the choice of lattice and there are optimal lattices that give minimum quantization error when combined with subtractive dithering. Here is a brief summary of the results of this chapter:

1. **Lattice Quantization Analysis.** We derive the statistical relationship between the lattice quantization noise and the quantizer input. The mathematical tool used is the extension of that used in the scalar case: Fourier series in multiple dimensions.

2. **Subtractive Dithering.** We introduce the notion of subtractive vector dithering and derive the necessary and sufficient conditions on the statistics of the dither vector to render the lattice quantization noise independent of the quantizer input. We then address the question of the best selection of the lattice that minimizes the reconstruction error. We give a necessary condition on the best

lattice. We also give efficient techniques to design suitable dither vectors.

3. **Nonsubtractive Dithering.** We examine the case of nonsubtractive vector dithering. We give necessary conditions for the second moment of the lattice quantization noise to be independent of the quantizer input statistics. We address the design of the dither vector that minimizes the reconstruction error.

4. **Optimal pre- and post-filtering of lattice quantizers.** The last section of the thesis deals with the question of optimal pre- and post-filtering of lattice quantizers. The filters are of MIMO type, and hence can be associated with filter banks. We clarify the relationship between this problem and the optimal biorthogonal subband coding problem. The main difference lies in the fact that, in the biorthogonal subband coding problem, one has scalar quantizers independently operating in the subbands. In contrast, in the pre- and post-filtering of lattice quantizers, the quantization of subbands is done jointly and there is no notion of bit allocation. It turns out, however, that if in the biorthogonal subband coding problem, equal number of bits are assigned to each of the subbands, then the two problems become mathematically equivalent to each other and therefore has the same solution.

1.4 Notations and Terminology

1. The notation $\tilde{X}(z)$ denotes the $z-$transform of $x^*(-n)$ where $*$ stands for complex conjugation. If $x(n)$ is real, then $\tilde{X}(z) = X(z^{-1})$. Notice that $\tilde{X}(z) = X^*(1/z^*)$, and the FT of $x^*(-n)$ is $X^*(e^{j\omega})$. In the matrix case, we have $\tilde{\mathbf{X}}(z) = \mathbf{X}^\dagger(1/z^*)$. The notation $\dagger$ in return means transpose and conjugate, that is $\mathbf{X}^\dagger = (\mathbf{X}^T)^*$.

2. The notation $X(z)|_{\downarrow M}$ denotes the $z-$transform of the downsampled sequence $x(Mn)$.

3. The notation $\delta(n)$ refers to an impulse sequence that is zero at all times except

the origin at which it is 1. That is, $\delta(0) = 1$, and $\delta(n) = 0,\ n \neq 0$.

4. **Nyquist(M) property.** A sequence $x(n)$ is said to be Nyquist(M) if $x(Mn) = \delta(n)$ or equivalently $X(z)|_{\downarrow M} = 1$. This can be rewritten in the form [Vai93]:

$$\sum_{k=0}^{M-1} X(zW^k) = M \tag{1.10}$$

where $W = e^{-j2\pi/M}$.

Chapter 2

Optimum Orthonormal Subband Coding

Optimization of filter banks for subband coding (SBC) of signals has been an active area of research [AM96, AL91, MM98, Vai98, VK98a]. Subband coding involves a linear transform part and a nonlinear quantization part as shown in Fig. 2.1. Block

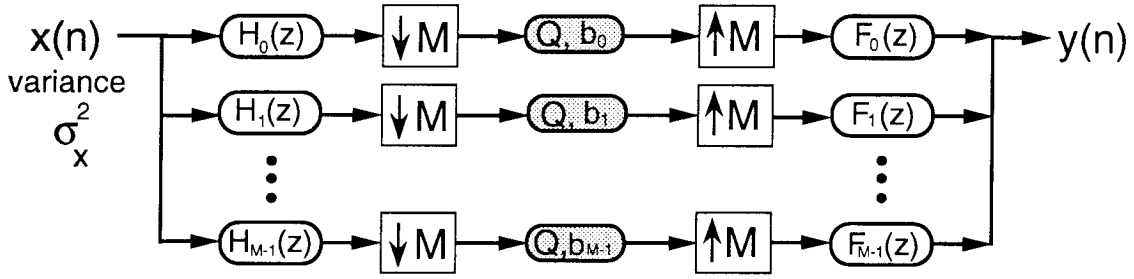


Fig. 2.1: Subband coding scheme with a uniform filter bank.

transform coding, overlapped transform coding, and wavelet-based coding are special forms of subband coding.

In this chapter we mainly deal with the optimization of orthonormal filter banks for subband coding. We first cover the uniform case and then present some of the first results for the nonuniform case. In the uniform case, optimal orthonormal filter banks with unconstrained filters are constructed in [Vai98] under the assumption that the quantizers operate at high bit rates. We show in this chapter that this construction continues to be optimal even in the case where the quantizers cannot be assumed to

have high bit rates. To do this, we first connect the problem of optimal orthonormal subband coding to that of principal component representation of signals. The latter is done via the so called principal component filter banks (PCFB). When the filter orders are unconstrained (ideal filter banks), PCFB's can always be constructed, and we show that they are optimal for orthonormal subband coding for all bit rates. In the FIR case, however, we present examples which show that the existence of PCFB's is not always guaranteed. Optimization of FIR filter banks for subband coding currently relies on nonlinear numerical optimization techniques. If one is willing to sacrifice optimality, then there are efficient ways of designing FIR filter banks. One such example is described in [MM98]. The technique utilizes the design of compaction filters and we will explain it in some detail at several points in this and the next chapter. The term principal component filter bank as used in [XB98] is not the same as the one we define in this chapter. What is meant by an FIR PCFB in [XB98] is actually an optimal FIR orthonormal filter bank that maximizes the coding gain under the high bit rate assumptions.

We then consider the problem of optimal nonuniform orthonormal subband coding. We extend the problem of principal component representation of signals to the nonuniform case. In contrast to the uniform case where there is single PCFB for a given input statistics, there are more than one nonuniform PCFB's, one for each ordering of the set of decimation ratios. However, we show that one of these PCFB's is optimal for nonuniform subband coding. We then address the question of optimal selection of the set of decimation ratios for a given number of channels. In the nonuniform case we are not concerned with order constraints, so the filters will be ideal.

The results of this chapter have been presented at various conferences [KV98a, KV98c, KV98b]. Our contributions in the biorthogonal subband coding case can be found in [VK98b, VK98a].

2.1 Uniform Case

Consider Fig. 2.1 again, where we show a subband coding scheme with a uniform filter bank as its transform part. Fig. 2.2 shows the polyphase representation for the analysis

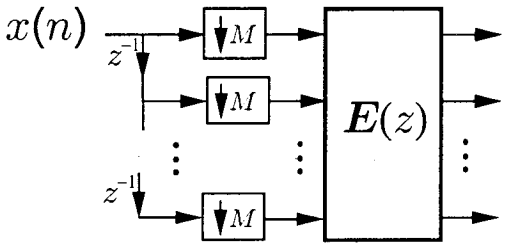


Fig. 2.2: Polyphase representation of the analysis part.

part [Vai93]. An orthonormal filter bank is one with a paraunitary $\mathbf{E}(z)$ which means that $\mathbf{E}(e^{j\omega})$ is unitary for each ω. In the special block transform coding case, we have $\mathbf{E}(z) = \mathbf{E}_0$, a constant matrix, and orthonormality is the same as $\mathbf{E}_0$ being unitary. In this case, the optimal orthogonal transform matrix is well known to be the Karhunen-Loeve transform. To show the optimality of KLT, one usually resorts to a quantization model that is valid only at high bit rates. However, a classic work by Huang and Schultheiss [HS63] contains a proof that KLT is optimal even if the quantizers cannot be assumed to have high bit rates. Inspired by this pioneering work, we provide a very simple proof of optimality of orthonormal filter banks under a general quantizer noise model without assuming high bit rates. Our simplified point of view enables us to state a very strong result that principal component filter banks are optimal for subband coding for all bit rates and bit allocation strategies.

We have been able to show that the problems of optimal representation of signals using principal component analysis and optimal orthonormal subband coding are fundamentally the same. It should be noted, however, that in general, biorthogonal filter banks perform better than orthonormal filter banks in subband coding [VK98a]. The only exception to this is the case of block transform coding where the KLT is orthonormal and no other biorthogonal transform can result in a better coding performance [Vai98].

2.1.1 Review of High Bit Rate Case

Let σ_x^2 be the original signal variance, $\sigma_{x_j}^2$ be the subband variances and $\sigma_{q_j}^2$ be the quantization noise variances in the subbands. If we assume that the quantizers operate at high bit rates, then one can model them by a simple relationship:

$$\sigma_{q_j}^2 = c2^{-2b_j}\sigma_{x_j}^2 \tag{2.1}$$

where b_j is the number of bits at which the quantizer in jth channel is operating. If the total bit budget is b bits per pixel, then we have

$$b = \sum_{j=0}^{M-1} b_j/M \tag{2.2}$$

Coding gain of a SBC scheme is defined to be the ratio of the quantization error when the input is quantized directly and the reconstruction error of the SBC scheme using the same bit budget b. In the high bit rate case, since we have analytical expressions for quantizer variances as in (2.1), we can optimally allocate the bits among the channels and finally obtain the following formula for coding gain [JN84]:

$$G_{coding} = \frac{\sigma_x^2}{(\prod_{j=0}^{M-1}\sigma_{x_j}^2)^{1/M}} \tag{2.3}$$

We have $\sigma_x^2 = \sum_{j=0}^{M-1}\sigma_{x_j}^2/M$ by orthonormality of the filter bank. Hence the coding gain in the high bit rate case is the ratio of arithmetic and geometric means of subband variances. In the transform coding case, by a well known inequality in linear algebra, this is maximized if and only if the transform is the KLT which diagonalizes the autocorrelation matrix of the input.

Block transform coding with $\mathbf{E}(z) = \mathbf{E}_0$ corresponds to subband coding with filters having orders less than the number of channels. Recently optimal orthonormal filter banks for subband coding with ideal filters have been constructed in [Vai98]. The construction is proven to be optimal under the assumption that the quantizers operate at high bit rates and therefore the coding gain expression (2.3) is valid. The optimal

paraunitary polyphase matrix $\mathbf{E}(e^{j\omega})$ is shown to have two properties [Vai98]:

1. It diagonalizes the power spectrum matrix of the input vector at each frequency (total decorrelation).

2. Its rows at each frequency are ordered in such a way that the corresponding diagonal elements are ordered as well (spectral majorization).

Loosely speaking, optimal $\mathbf{E}(e^{j\omega})$ for a given ω is the Karhunen-Loeve transform matrix corresponding to the input power spectrum matrix evaluated at the frequency ω. In the biorthogonal case, it is shown in [VK98a] that the above orthonormal solution can be improved by further processing of subband signals. This final stage consists of half-whitening of uncorrelated subband signals.

2.1.2 Arbitrary Bit Rate Case

A characterizing property of an orthonormal transform is that it preserves the total energy of its input. Hence orthonormal transforms are also referred as lossless transforms. This is the major property we use to prove the optimality of an orthonormal filter bank for subband coding at arbitrary bit rates. Consider Fig. 2.1 again. Let $\sigma_x^2, \sigma_{x_j}^2$, and $\sigma_{q_j}^2$ be variances of the original input, subband signals, and the quantization noise signals as defined before. Let $\mathcal{E}$ denote the reconstruction error of this subband coding scheme. Then, by the orthonormality of the transform, we have:

$$\mathcal{E} = \sum_{j=0}^{M-1} \sigma_{q_j}^2 / M \tag{2.4}$$

A filter bank is said to be optimum for subband coding if the reconstruction error $\mathcal{E}$ is minimized for a given bit budget, b bits per input sample as in (2.2). At high bit rates, optimality is the same as maximization of the expression in (2.3). In general, however, (2.3) does not represent the objective.

A More General Quantizer Model

We will model all the quantizers in Fig. 2.1 by a single quantization function $f(.)$ such that

$$\sigma_{q_j}^2 = f(b_j)\sigma_{x_j}^2 \tag{2.5}$$

The important benefit of this model is that **we do not assume high bit rates**. However, we do assume that all the channels have the same quantizer function $f(.)$, which in general need not be true. In general, the quantizer functions $f_j(.)$ depend on the statistics of the jth channel signal, which in turn depend on the linear transformation. If, however, the original input signal is Gaussian, then regardless of which transform is used, the subband channels are all Gaussian as well, and hence can be modeled by a single quantizer function $f(.)$. We do not assume that the noise signals are white or uncorrelated with each other. The traditional high bit rate quantizer model can now be considered as a special case of our model with $f(b_j) = c2^{-2b_j}$. With our quantizer model, using the relations (2.4) and (2.5) the reconstruction error of the subband coding scheme is

$$\mathcal{E} = \sum_{j=0}^{M-1} f(b_j)\sigma_{x_j}^2/M \tag{2.6}$$

2.1.3 Principal Component Filter Banks (PCFB)

Let us consider for a moment a different but a closely related problem that deals with multiresolution representation of signals. Consider Fig. 2.3. If we keep the first $P < M$

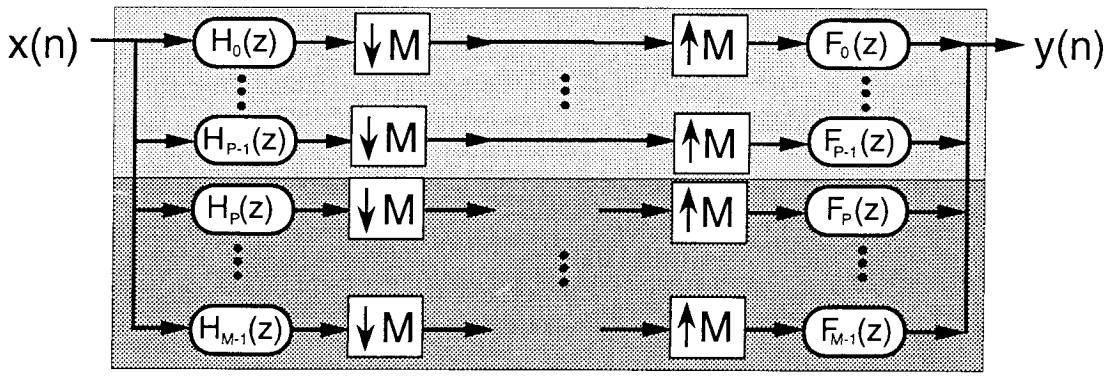


Fig. 2.3: Pertaining to the discussion of principal component filter banks.

of the subband signals without quantizing, and if we drop the other subbands, there will be a corresponding reconstruction error. Intuitively, the minimization of this error is the same as putting most of the signal information into the first P channels. If there is a single filter bank that minimizes this error for each $P = 1, \ldots, M$, we call it a **principal component filter bank**. It provides us an optimal multiresolution representation of the signal as P increases from 1 to M. For $P = M$, the reconstruction error is zero if the filter bank is perfect reconstruction (that is, biorthogonal). It turns out that, if we allow ideal filters, there always exists a principal component filter bank [TG95] for any given input process, and it is orthonormal. Furthermore, it is the same filter bank as the optimal orthonormal filter bank for subband coding as derived in [Vai98] under high bit rate quantizer models. Is this a coincidence? The relationship between the two problems are clarified in the next section under our more general modeling of quantizers.

2.1.4 Optimality of PCFB's Under Arbitrary Bit Rates

Consider Fig. 2.3 again where the first $P < M$ of the subbands are kept. If the filter bank is orthonormal, then the reconstruction error is the sum of variances of dropped subbands, that is:

$$\mathcal{E} = \sum_{j=P}^{M-1} \sigma_{x_j}^2 / M \tag{2.7}$$

Notice the similarity between (2.6) and (2.7). The latter can be seen as a special case of the former when $f(b_j) = 0,\ j = 0, \ldots, P-1$ (infinite precision) and $f(b_j) = 1,\ j = P, \ldots, M-1$ (zero bit assignment). By the lossless property, the sum of all subband variances is constant as noted before. Hence the error in (2.7) is minimized if $\sum_{j=0}^{P-1} \sigma_{x_j}^2 / M$ is maximized. A principal component filter bank as defined in the previous section maximizes the partial sum $\sum_{j=0}^{P-1} \sigma_{x_j}^2 / M$ for each P. We are now ready to state and prove our main result:

Theorem 1. *With a general quantizer model described in Sec. 2.1.2, a principal component filter bank, if it exists, minimizes the reconstruction error of an orthonormal subband coder for all bit budgets and bit allocation strategies.* ◇

Proof. Without loss of generality assume that $f(b_0) \leq \ldots \leq f(b_{M-1})$. By simple algebra, we can write the error in (2.6) as

$$\mathcal{E} = \sum_{j=1}^{M-1} (f(b_{j-1}) - f(b_j)) \sum_{i=0}^{j-1} \sigma_{x_i}^2/M + f(b_{M-1}) \sum_{i=0}^{M-1} \sigma_{x_i}^2/M \tag{2.8}$$

By assumption, $f(b_{j-1}) - f(b_j) \leq 0$. The last term is fixed and equal to $f(b_{M-1})\sigma_x^2$. Hence $\mathcal{E}$ is minimized if the partial sums $\sum_{i=0}^{j-1} \sigma_{x_i}^2$ are maximized for each j. This is the case if the filter bank is a principal component filter bank (PCFB). ■

Remarks.

1. Notice that optimal bit allocation did not enter the discussion. A PCFB minimizes $\mathcal{E}$ for any bit allocation, in particular, for the optimal bit allocation. There are many interesting results on the optimum allocation of bits. We refer to the authorative work [SG88] for a general treatment of the topic. As long as we can model the quantizers with one underlying quantization function, the optimization of the orthonormal filter bank is decoupled from the optimization of bit allocation in subbands. *Even if the bit allocation is not optimal, a PCFB continues to be optimal.*

2. From the definition, the subband variances of a PCFB are in decreasing order: $\sigma_{x_0}^2 \geq \sigma_{x_1}^2 \geq \ldots \geq \sigma_{x_{M-1}}^2$. This is because, if this is not true, then one can get a new filter bank by reordering the filters such that the variances are ordered. The new filter bank will have a higher sum $\sum_{j=0}^{P-1} \sigma_{x_j}^2$ for at least one P contradicting that the original filter bank was a PCFB.

3. Given the set of ordered variances $\sigma_{x_0}^2 \geq \sigma_{x_1}^2 \geq \ldots \geq \sigma_{x_{M-1}}^2$ of a PCFB, one can do optimal bit allocation using the quantizer function $f(.)$. Let $b_j^*, j = 0, \ldots, M-1$ denote this optimum allocation. If the inequalities between subband variances are strict, i.e., $\sigma_{x_0}^2 > \sigma_{x_1}^2 > \ldots > \sigma_{x_{M-1}}^2$, then we must have $f(b_0^*) \leq \ldots \leq f(b_{M-1}^*)$.

To see this, assume $f(b_{i-1}^*) > f(b_i^*)$ for some i. Then by interchanging b_{i-1}^* and b_i^*, we see that the total error $\mathcal{E} = \sum_{j=0}^{M-1} f(b_j^*)\sigma_{x_j}^2/M$ is reduced while maintaining the total bit budget, contradicting the optimality of the bit allocation. If, however, for any i, $\sigma_{x_{i-1}}^2 = \sigma_{x_i}^2$, then one can interchange the bits to guarantee $f(b_{i-1}^*) \leq f(b_i^*)$ without affecting the total error.

4. As an artificial special case, let $P < M$ be fixed and let $f(b_j) = 0$, $j = 0, \ldots, P-1$, and $f(b_j) = 1$, $j = P, \ldots, M-1$. Then we have to maximize $\sum_{j=0}^{P-1} \sigma_{x_j}^2$ for that P. If we repeat the process for each $P = 1, \ldots, M-1$, and if there exists one single solution for all of them, then the solution is a PCFB!

5. As another special case, consider $f(b_j) = c$, a constant, for all j. This corresponds to equal bit allocation strategy. In this case we have $\mathcal{E} = c\,\sigma_x^2$, which is independent of the transformation. Hence, orthonormal filter banks do not yield any coding advantage if equal bit allocation strategy is used.

6. Finally consider the case $f(b_0) < f(b_1) < \ldots < f(b_{M-1})$. In this case, all $f(b_{j-1}) - f(b_j) < 0$, and therefore the partial sums $\sum_{i=0}^{j-1} \sigma_{x_i}^2$ must be maximized for all j. Hence a PCFB, if exists, is not only sufficient but also necessary for optimality.

Underlying Mathematical Theory

The result that we have just presented is yet another application of a topic in applied mathematics called **majorization** [MO79]. A set of M numbers a_j, $j = 0, \ldots, M-1$ is said to majorize another set of M numbers b_j, $j = 0, \ldots, M-1$, if

$$\sum_{j=0}^{P-1} a_j \geq \sum_{j=0}^{P-1} b_j, \quad P = 1, \ldots, M, \tag{2.9}$$

with equality if $P = M$. We see that a PCFB is the one that produces a set of subband variances that majorizes sets of subband variances obtained by all other orthonormal filter banks. In the special high bit rate case, the proof that a PCFB is optimal follows

directly from the fact that

$$\prod_{j=0}^{M-1} a_j \leq \prod_{j=0}^{M-1} b_j \tag{2.10}$$

whenever the inequalities in (2.9) hold with equality if $P = M$ [HJ85]. There are many other applications of majorization. The interested reader is referred to [MO79] for a beautiful treatment of the subject and to [SS94] for more recent applications.

2.2 Uniform FIR Case

The result developed in the previous section asserts that for a given input, whenever one orthonormal filter bank produces a set of subband variances that majorizes sets of subband variances of all other orthonormal filter banks, it is optimal for orthonormal subband coding for that input under all bit rates and bit allocation strategies. We have not yet put any constraints on the filters. Now, we can think of a class of orthonormal filter banks that is characterized by practical constraints such as finite filter lengths. The simplest case is the block transforming case we have considered at the beginning. In this case, the autocorrelation matrices of subband and original input signals are related via a unitary matrix. It is a well known fact in linear algebra that the eigenvalues of an autocorrelation matrix majorizes its diagonal elements. This gives us the simplified proof that KLT is optimal among the class of block transforms. This is because, with KLT, eigenvalues become the subband variances. As before, the optimality is independent of the bit rates involved and bit allocation strategies.

So, in the two extreme cases where (1) filters can be ideal and (2) filters are constrained to be of the same length as the number of channels, we know that a PCFB exists and therefore optimal for orthonormal subband coding. What happens in the intermediate case where the filter lengths are finite but larger than the number of channels? This case turns out to be very difficult to analyze as confirmed by several researchers [MM98, Uns93a, XB98] who devised numerical techniques for suboptimal solutions. One approach for a suboptimal solution is to design an optimal energy compaction filter [KV98e] that pushes most of the signal energy into the first channel, and

then complete the filter bank in some optimal fashion [MM98]. Although this technique was claimed to find optimal solution in [MM98], we show in this section that this is not the case. The reason for its suboptimality is that it is a loss of generality to design the first filter to be a compaction filter, as there may not even exist a PCFB! Another important approach is to develop structures that are efficient to implement [Mal92] while hopefully not being too much away from the optimal solution. At present, there is no robust algorithm that converges to an optimal FIR orthonormal filter bank. A good strategy for large filter lengths is to try to approximate the optimal ideal filter banks.

The difficulty in FIR case starts with the exact definition of the class of filter banks in which an optimal solution is searched for. There are two equally plausible choices: those with a finite degree and those with a finite order. Consider the subband coding scheme in Fig. 2.1 and the polyphase representation of its analysis part in Fig. 2.2. Assume

$$\mathbf{E}(z) = \sum_{n=0}^{K} \mathbf{E}_n z^{-n} \tag{2.11}$$

Here $\mathbf{E}_n$'s are $M \times M$ constant matrices with $\mathbf{E}_K \neq \mathbf{0}$. The **order** of $\mathbf{E}(z)$ is said to be K. Note that filter lengths can be as high as $M(K+1)$. The notion of **degree** is different from the notion of order and it is defined to be the minimum number of delay elements to implement the MIMO system $\mathbf{E}(z)$. If μ denotes degree, then in general, $\mu \geq K$. Notice that this ambiguity does not arise in block transform and ideal filters cases. In the former, both degree and order are zero, while in the latter they are infinite or undefined. The orthonormality of the filter bank is equivalent to

$$\mathbf{E}^{\dagger}(e^{j\omega})\mathbf{E}(e^{j\omega}) = \mathbf{I}, \ \forall\omega \tag{2.12}$$

In this case $\mathbf{E}(z)$ is also said to be paraunitary and it is well known that $\mathbf{E}(z)$ can be factored as [Vai93]

$$\mathbf{E}(z) = \mathbf{U}\mathbf{V}_1(z)\mathbf{V}_2(z)\ldots\mathbf{V}_\mu(z) \tag{2.13}$$

where $\mathbf{U}$ is unitary,

$$\mathbf{V}_n(z) = \mathbf{I} - \mathbf{v}_n\mathbf{v}_n^\dagger + z^{-1}\mathbf{v}_n\mathbf{v}_n^\dagger,\ n = 1, \ldots, \mu, \tag{2.14}$$

and $\mathbf{v}_n$'s have unit norm, i.e., $\mathbf{v}_n^\dagger\mathbf{v}_n = 1$. (see Fig. 2.4). Conversely any $\mathbf{E}(z)$ of the

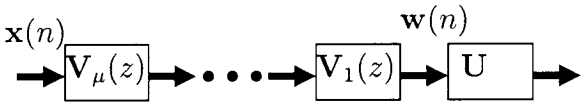


Fig. 2.4: Householder factorization of $\mathbf{E}(z)$.

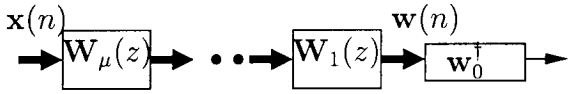


Fig. 2.5: Householder factorization of $\mathbf{e}_0^\dagger(z)$.

form (2.13) is FIR orthonormal of degree μ as long as $\mathbf{U}$ is unitary and $\mathbf{v}_n$'s have unit norm.

If we are interested in only one filter, say $H_0(z)$ of the filter bank as in the case of compaction problem, we need to consider only the corresponding row $\mathbf{e}_0^\dagger(z)$ of $\mathbf{E}(z)$ whose elements are the polyphase components of $H_0(z)$. Let ν be the degree of $\mathbf{e}_0^\dagger(z)$. Then similar to (2.13) we can write

$$\mathbf{e}_0^\dagger(z) = \mathbf{w}_0^\dagger\mathbf{W}_1(z)\mathbf{W}_2(z)\ldots\mathbf{W}_\nu(z) \tag{2.15}$$

where $\mathbf{W}_n(z) = \mathbf{I} - \mathbf{w}_n\mathbf{w}_n^\dagger + z^{-1}\mathbf{w}_n\mathbf{w}_n^\dagger,\ n = 1, \ldots, \nu$, and the vectors $\mathbf{w}_n,\ n = 0, \ldots, \nu$ have unit norm [Vai93] (See Fig. 2.5).

2.2.1 On Uniqueness of the Factorizations

The factorization of $M \times M$ polyphase matrix $\mathbf{E}(z)$ of a given degree μ is in general not unique and $K \leq \mu$. The factorization of $\mathbf{e}_0^\dagger(z)$ of degree ν is on the other hand unique. The order of $\mathbf{e}_0^\dagger(z)$ is equal to its degree ν [Vai93]. This implies the following: $\mathbf{w}_n^\dagger\mathbf{w}_{n+1} \neq 0, n = 0, \ldots, \nu - 1$.

Fact 1. If in the factorization of $\mathbf{E}(z)$ in (2.13), the vectors $\mathbf{v}_n$ turn out to be such that $\mathbf{v}_n^\dagger \mathbf{v}_{n+1} \neq 0, n = 1, \ldots, \mu - 1$, then we have $K = \mu$ and the factorization is unique. Otherwise, $K < \mu$ and the factorization is not unique.

Proof. From (2.13), we can write the highest possible coefficient

$$\mathbf{E}_\mu = \mathbf{U}\mathbf{v}_1\mathbf{v}_1^\dagger\mathbf{v}_2\mathbf{v}_2^\dagger \ldots \mathbf{v}_\mu\mathbf{v}_\mu^\dagger \tag{2.16}$$

Since $\mathbf{v}_n^\dagger \mathbf{v}_{n+1} \neq 0$, and since $\mathbf{U}$ is nonsingular, we conclude that $\mathbf{E}_\mu \neq 0$ and therefore $K = \mu$. The ith row of $\mathbf{E}(z)$ is $\mathbf{e}_i^\dagger(z) = \mathbf{u}_i^\dagger \mathbf{V}_1(z)\mathbf{V}_2(z)\ldots\mathbf{V}_\mu(z)$. There exists at least one index i, say $i = 0$, such that $\mathbf{u}_0^\dagger \mathbf{v}_1 \neq 0$ (otherwise $\mathbf{U}$ has to be singular). Hence the degree of $\mathbf{e}_0^\dagger(z)$ is μ. This implies that $\mathbf{V}_n(z)$'s are unique. Since $\mathbf{U} = \mathbf{E}(1)$ is unique, we conclude that the factorization (2.13) is indeed unique. If on the other hand we have $\mathbf{v}_n^\dagger \mathbf{v}_{n+1} = 0$ for some n, then it can be shown that

$$\mathbf{E}_0 = (\mathbf{I} - \mathbf{v}_1\mathbf{v}_1^\dagger)(\mathbf{I} - \mathbf{v}_2\mathbf{v}_2^\dagger)\ldots(\mathbf{I} - \mathbf{v}_\mu\mathbf{v}_\mu^\dagger) \tag{2.17}$$

has rank less than $M - 1$. This implies that the rank reduction [Vai93] can start with more than one possible vector $\mathbf{v}_\mu$, implying that the factorization is not unique. ■

Now, returning to $\mathbf{E}(z)$ of degree μ and its first row $\mathbf{e}_0^\dagger(z)$ of degree ν, assume that $\mu = \nu$. Then by the uniqueness of the representation (2.15) it follows that $\mathbf{V}_n(z) = \mathbf{W}_n(z)$, $n = 1, \ldots, \mu$, and $\mathbf{w}_0 = \mathbf{u}_0$ where $\mathbf{u}_0^\dagger$ is the first row of $\mathbf{U}$. Therefore, all the other filters $H_i(z), i = 1, \ldots, M - 1$ can be determined by the remaining $M - 1$ rows of the unitary matrix $\mathbf{U}$. This leads to the following design algorithm for signal-adapted FIR orthonormal filter banks originally proposed by Moulin et al. [MM98]:

1. Design the first filter $H_0(z)$ to be a compaction filter [KV98e].

2. Factorize the polyphase vector $\mathbf{e}_0^\dagger(z)$ of $H_0(z)$ as

$$\mathbf{e}_0^\dagger(z) = \mathbf{w}_0^\dagger \mathbf{W}_1(z)\mathbf{W}_2(z)\ldots\mathbf{W}_\mu(z) \tag{2.18}$$

Let $\mathbf{V}_n(z) = \mathbf{W}_n(z)$, $n = 1, \ldots, \mu$.

3. Choose $\mathbf{U}$ to be the KLT for its input vector. The first row of the KLT is necessarily $\mathbf{w}_0^\dagger$. If it is not, one can increase the compaction gain, violating the optimal compaction property of $H_0(z)$.

The authors of [MM98] use the argument that if one designs a principal component filter bank (PCFB), then it maximizes the coding gain. The first filter of a PCFB has to be a compaction filter. Hence the above algorithm should be optimum. They assume implicitly that a PCFB exists. If the ideal filters are allowed, then a PCFB does exist and it maximizes the coding gain [TG95, Vai98]. Similarly, if the filter orders are less than the number of channels, then the KLT achieves the maximum coding gain and it is a PCFB. We show later in this section that in the intermediate case, there does not always exist a PCFB. Hence the above algorithm is in general suboptimum. Nevertheless, as we show by some examples in Sec. 2.2.4, the suboptimality is not significant for practical signals. Since the design of FIR compaction filters is well studied [KV98e, Mou95, TV98] and there exist very efficient algorithms like the **window method** that we describe in Chapter 3, we see that the above method is very efficient for the design of signal-adapted FIR orthonormal filter banks.

2.2.2 FIR Coding and Compaction Problems

Since we assume that $x(n)$ is WSS, the vector process $\mathbf{x}(n)$ that is the input of $\mathbf{E}(z)$ in Fig. 2.2 is WSS with power spectral density (psd) matrix $\mathbf{S_{xx}}(e^{j\omega})$. Using the fact that $\mathbf{E}(z)$ is FIR paraunitary, with high-bit rate assumptions on the quantization noise sources, and with optimal bit allocation, the reconstruction error is $\mathcal{E} = c2^{-2b}\phi^{1/M}$ where [Vai98]

$$\phi = \prod_{i=0}^{M-1} \sigma_{x_i}^2 = \prod_{i=0}^{M-1} \int_{-\pi}^{\pi} \left[\mathbf{E}^\dagger(e^{j\omega})\mathbf{S_{xx}}(e^{j\omega})\mathbf{E}(e^{j\omega})\right]_{ii} \frac{d\omega}{2\pi} \tag{2.19}$$

Here $\sigma_{x_i}^2$ is the variance of the ith subband. Let $\mathcal{O}_\mu$ denote the class of $M \times M$ FIR orthonormal polyphase matrices with degree less than or equal to μ. The **coding**

problem with the high bit rate assumptions is the following:

$$\min_{\mathbf{E}(z)\in\mathcal{O}_\mu} \prod_{i=0}^{M-1} \int_{-\pi}^{\pi} \left[\mathbf{E}^\dagger(e^{j\omega})\mathbf{S_{xx}}(e^{j\omega})\mathbf{E}(e^{j\omega})\right]_{ii} \frac{d\omega}{2\pi} \tag{2.20}$$

The energy compaction problem, on the other hand, is concerned with making one of the subband variances of an orthonormal filter bank as large as possible. If the original signal is WSS, then the compaction gain is defined as $G_{comp} = \max_i (\sigma_{x_i}^2)/\sigma_x^2$. Let $\mathcal{Q}_\mu$ denote the class of $1 \times M$ FIR orthonormal polyphase vectors of degree less than or equal to μ. The **compaction problem** is the following:

$$\max_{\mathbf{e_0}(z)\in\mathcal{Q}_\mu} \int_{-\pi}^{\pi} \mathbf{e}_0^\dagger(e^{j\omega})\mathbf{S_{xx}}(e^{j\omega})\mathbf{e}_0(e^{j\omega}) \frac{d\omega}{2\pi} \tag{2.21}$$

Considering Fig. 2.4, the objectives can be written as

$$\min_{\mathbf{v}_n,\mathbf{U}} \prod_{i=0}^{M-1} [\mathbf{U}\mathbf{R_{ww}}(0)\mathbf{U}^\dagger]_{ii} \quad \text{(coding)} \tag{2.22}$$

$$\max_{\mathbf{v}_n,\mathbf{u}_0} \mathbf{u}_0^\dagger \mathbf{R_{ww}}(0)\mathbf{u}_0 \quad \text{(compaction)} \tag{2.23}$$

where $\mathbf{R_{ww}}(0)$ is the autocorrelation matrix of $\mathbf{w}(n)$. In the coding problem, $\mathbf{U}$ has to be the KLT for $\mathbf{w}(n)$ and in the compaction problem $\mathbf{u}_0$ has to be the unit-norm eigenvector of $\mathbf{R_{ww}}(0)$ corresponding to the maximum eigenvalue. Let λ_i's be the eigenvalues of $\mathbf{R_{ww}}(0)$. Hence one can rewrite the problems as:

$$\min_{\mathbf{v}_n} \prod_{i=0}^{M-1} \lambda_i \quad \text{(coding)}, \quad \max_{\mathbf{v}_n} \max_i \lambda_i \quad \text{(compaction)} \tag{2.24}$$

Hence both problems are parameterized by μ unit-norm vectors of length M. The total number of free parameters is therefore $\mu(M-1)$. If $\mu = 0$, there is nothing to optimize. In this case $\mathbf{E}(z) = \mathbf{E}_0 = \mathbf{U}$ is the KLT for the input vector $\mathbf{x}(n)$ in the coding problem and $\mathbf{u}_0^\dagger$ is the first row of the KLT in the compaction problem. The matrix $\mathbf{U}$ diagonalizes $\mathbf{R_{xx}}(0)$, the $M \times M$ autocorrelation matrix of the input. The solution for the case where the filter orders are unconstrained has recently been estab-

lished. We will refer to this case as $\mu = \infty$, although the degree is formally undefined because the filters are not causal. The optimum solution $\mathbf{E}(e^{j\omega})$ that maximizes the coding gain, diagonalizes $\mathbf{S_{xx}}(e^{j\omega})$ at each frequency. This in particular implies the diagonalization of the autocorrelation matrix $\mathbf{R_{xx}}(0)$ (which was both necessary and sufficient condition for the transform coding case). Diagonalization of the psd matrix at each frequency, however, is not sufficient for $\mathbf{E}(e^{j\omega})$ to maximize the coding gain [Vai98]. There should be an additional ordering of the eigenvalues of the psd matrix at each frequency (spectral majorization) [Vai98]. If $x(n)$ is WSS, then these eigenvalues are $S_{xx}(e^{j(\omega+i2\pi/M)})$, $i = 0, \ldots, M-1$. For the two-channel case and for a restricted class of input psd, we show in Chapter 3 that if μ is the degree of the optimum FIR filter bank, then $\mathbf{S_{xx}}(e^{j\omega})$ should be decorrelated and majorized only at $\lceil \mu/2 \rceil$ discrete frequencies. In Chapter 3 we show how to find those frequencies.

2.2.3 Existence of FIR PCFB's: A Counter Example

Let the input process be AR(1) with the correlation coefficient of $\rho = 0.9$. Let the number of channels be $M = 3$ and $\mu = K = 1$. Assume that the filter orders are less than or equal to $N = 4$. Note that these are the smallest numbers for which we can expect to have a counter example. This is because the coding and compaction problems are the same if either $M = 2$ or $N < M$ [KV98e]. Now, since the maximum filter order is 4, we can write $\mathbf{v}_1 = [\cos(\alpha)\ \sin(\alpha)\ 0]^T$. Hence the two problems can be formulated by one single parameter α. Hence we can plot the coding and compaction gains versus α as in Fig. 2.6 where we kept the range of α from 0 to $\pi/2$. This is because the plots are symmetric with respect to both 0 and $\pi/2$. From the plot we see that the two problems have different answers. The value of α that maximizes the coding gain is $\alpha_{coding} = 0.1507\pi$ whereas $\alpha_{comp} = 0.1695\pi$ maximizes the compaction gain. For these choices of α, the coding gains are $G_{coding} = 3.2176$ and $G_{coding} = 3.2052$ respectively, and the compaction gains are $G_{comp} = 2.7672$ and $G_{comp} = 2.7682$. Now we can conclude the following fact:

Fact 2. In general, there does not exist an FIR $M-$channel PCFB for finite nonzero

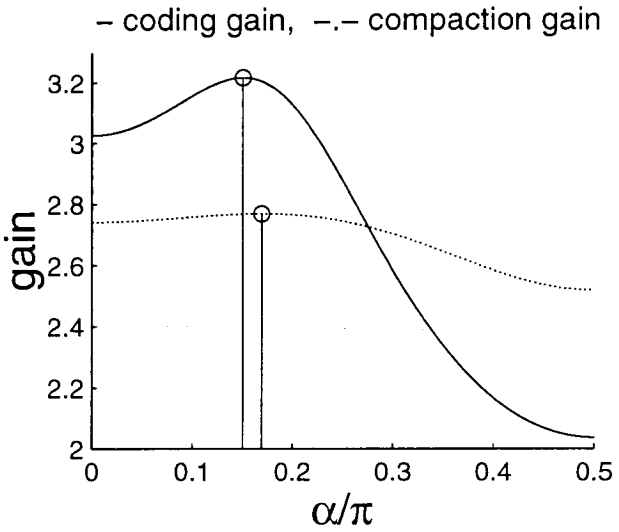


Fig. 2.6: Coding and compaction gains versus the parameter α. The two plots have maxima at different values of α.

degree μ.

Proof. Assume on the contrary that there always exists a PCFB $\mathbf{E}^p(z)$ of degree $0 < \mu < \infty$. Then for all $\mathbf{E}(z) \in \mathcal{O}_\mu$, $\sum_{i=0}^{P-1} \sigma_{x_i}^2$ is maximized by $\mathbf{E}^p(z)$, for each $P = 1, \ldots, M$. This implies two things: $\sigma_{x_0}^2$ is maximized by $\mathbf{E}^p(z)$ (optimum compaction gain), and $\prod_{i=0}^{M-1} \sigma_{x_i}^2$ is minimized by $\mathbf{E}^p(z)$ (optimum coding gain). The first one is by definition ($P = 0$), while the second one is due Theorem 1. So, if a PCFB exists, it solves both optimization problems. Since we have the above counter example as well as other examples in Sec. 2.2.4 that show that there is no single filter bank that achieves both the maximum compaction and coding gains, we conclude that a PCFB of a given degree does not always exist. ■

In the above example, among the class of orthonormal filter banks with $\mu = 1$ and the maximum filter order $N = 4$, there does not exist a PCFB. If it existed, then it would have achieved both the maximum compaction and coding gains. From the plot in Fig. 2.6, we see that there is no value of α for which both gains are maximized.

We have just shown that an FIR PCFB of a given degree μ does not always exist. The following example shows that an FIR PCFB of a given order K does not always exist either. These results help us understand why it is difficult to find analytical

solutions for optimal $M-$channel uniform FIR orthonormal filter banks.

Example 1. Let $M = 3$, and $K = 1$ (LOT case). If a PCFB exists, then it solves the following two problems: maximize $\sigma_{x_0}^2$, and maximize $\sigma_{x_0}^2 + \sigma_{x_1}^2$. The second problem is the same as the minimization problem: minimize $\sigma_{x_2}^2$. This is because, by orthonormality, $\sigma_{x_0}^2 + \sigma_{x_1}^2 + \sigma_{x_2}^2 = 3\sigma_x^2$. Both two problems can be seen as the problem of designing optimal FIR compaction filters [KV98e, TV98]. We have used the numerical technique developed in [TV98] which is guaranteed to converge to optimal solutions. The solution to the maximization problem consists of a set of filter coefficients for the first filter $H_0(z)$. These correspond to all possible spectral factors of a single magnitude-squared response. Similarly, the minimization of $\sigma_{x_2}^2$ leads to a set of coefficients for the third filter $H_2(z)$. If it turns out that no combination of filters from the two sets of solutions can possibly belong to a single orthonormal filter bank, then this is a proof that a PCFB of order $K = 1$ does not exist. We next show that this is the case. We can uniquely factorize the polyphase vector of each solution $H_0(z)$ into Householder factors: $\mathbf{e}_0^\dagger(z) = \mathbf{u}_0^\dagger V_1(z)V_2(z)$. Similarly for each $H_2(z)$ we have $\mathbf{e}_2^\dagger(z) = \mathbf{u}_2^\dagger W_1(z)W_2(z)$. The Householder factors are in the form $V(z) = I - \mathbf{v}\mathbf{v}^\dagger + z^{-1}\mathbf{v}\mathbf{v}^\dagger$ so that $V(1) = I$. Hence $\mathbf{e}_0^\dagger(1) = \mathbf{u}_0^\dagger$ and $\mathbf{e}_2^\dagger(1) = \mathbf{u}_2^\dagger$. If there was a single orthonormal filter bank, we would have had $\mathbf{u}_0^\dagger\mathbf{u}_2 = 0$ for at least one combination of solutions. For an AR(1) process with $\rho = 0.9$ we have explicitly found that this was not the case. Hence we conclude that, for this example, there does not exist a PCFB of order $K = 1$.

2.2.4 Efficiency of the Suboptimum Design

In Sec. 2.2.1, we have outlined an algorithm proposed by Moulin et al. [MM98] to design signal-adapted FIR orthonormal filter banks. In the algorithm, the first filter of the filter bank is constrained to be an optimum compaction filter. In the previous section we have shown an example where this constraint resulted in loss of coding gain. Another issue with this algorithm is the fact that the optimum compaction filter $H_0(z)$ is not uniquely determined from its magnitude square (or the product filter) $|H_0(e^{j\omega})|^2$. Since the latter can be spectrally factorized in many ways, we see that one

spectral factor may give better coding gain than the others although they all have the same compaction gain. This indeed turns out to be the case as we show in Example 3. In that example, we show also that even if one uses the compaction filter that has the best phase response (best spectral factor of $|H_0(e^{j\omega})|^2$), one can still increase the coding gain by brute-force optimization of the filter bank. We want to remark that the coding gain loss due to constraining the first filter to be optimum compaction filter is not significant for most of the practical signals we have considered. Below are some examples that confirm this observation.

Example 2. Let us consider the counter example of the previous section. Let the input be MA(1) instead of AR(1) with arbitrary correlation. Then one can verify by explicitly plotting the coding and compaction gains versus the parameter α that both achieve the maximum at the same value of α. This means that the best coding gain is achieved by designing a compaction filter first. This determines $\mathbf{v}_1$ and the first row of $\mathbf{U}$. The best filter bank that maximizes the coding gain is then obtained by using the KLT for the output of $\mathbf{V}_1(z)$. In the previous section, the difference in the coding gains was very small; for this example it is identically zero.

Example 3. Let the input be MA(1) with $r(0) = 1$ and $r(1) = 0.3$. Let $M = 3$ and $\mu = 5$ so that the maximum filter order, $N \leq 17$. The best compaction gain is $G_{comp} = 1.4920$ achieved by the best compaction filter magnitude response. There are eight possible phase responses for $H_0(e^{j\omega})$ that yield the same magnitude response $|H_0(e^{j\omega})|^2$ (assuming real coefficients). Among them there is one filter that achieves the maximum coding gain of $G_{coding} = 1.0944$. The minimum phase filter has the coding gain of $G_{coding} = 1.0653$ which is worse. By brute-force optimization, one can find a filter bank that has the coding gain of $G_{coding} = 1.0951$. This has a compaction gain of $G_{comp} = 1.4910$, slightly worse than the optimum. Hence this is an example where the phase response of the compaction filter $H_0(e^{j\omega})$ does affect the coding gain and even with a best phase, the coding gain is not the maximum achievable. On the other hand the numerical differences are not significant at all.

Example 4. Let the input be an AR(12) process. Let $M = 8$ and the degree $\mu = 5$ so that the maximum filter order $N \leq 47$. This is the example where we obtained

the most discrepancy between the two solutions. The coding gain for the suboptimum method of Moulin et al. is $G_{coding} = 5.3948$. By brute-force optimization of vectors $\mathbf{v}_n$, we find that we can achieve a coding gain of $G_{coding} = 5.9642$. The previous solution has the maximum compaction gain $G_{comp} = 5.9190$ while the latter solution has $G_{comp} = 5.0228$.

2.2.5 The Two-channel Case

In the special two-channel case, a PCFB is the one that maximizes $\sigma_{x_0}^2$. Hence an FIR PCFB can always be constructed by designing the first filter to be an optimal FIR compaction filter. The second filter is determined from the first. Hence, in the two-channel case, an FIR PCFB always exists for any input and it is optimal for orthonormal subband coding for all bit rates and bit allocation strategies. Furthermore, they can be analytically constructed for some special class of input signals as we show in Chapter 3.

2.3 Nonuniform Case

In the implementation of wavelet-based coders, one uses a dyadic tree-structured filter

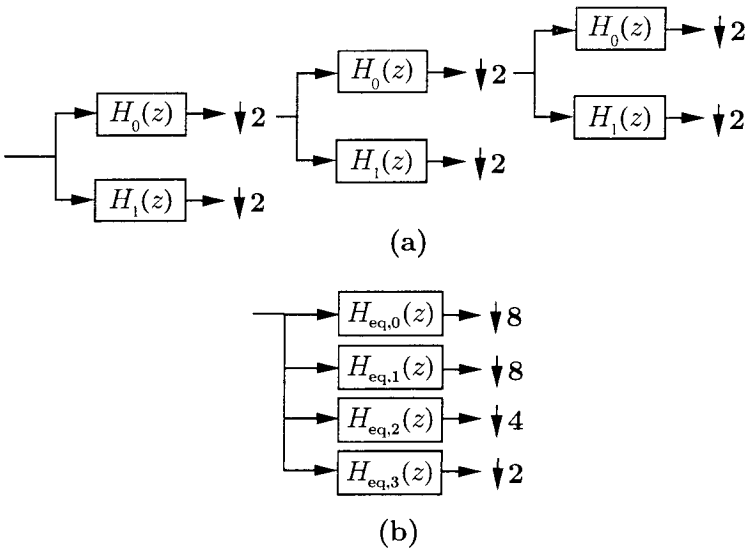


Fig. 2.7: (a) A three level wavelet decomposition, (b) equivalent nonuniform filter bank.

bank. This is equivalent to a nonuniform filter bank with decimation ratios that are powers of two. As an example, a three level wavelet decomposition is equivalent to a nonuniform filter bank with decimation ratios $\{8, 8, 4, 2\}$ as illustrated in Fig. 2.7. More generally, there are wavelet-packet based coders which incorporate arbitrary tree-structured filter banks, an example of which is shown in Fig. 2.8a. For simplicity, we

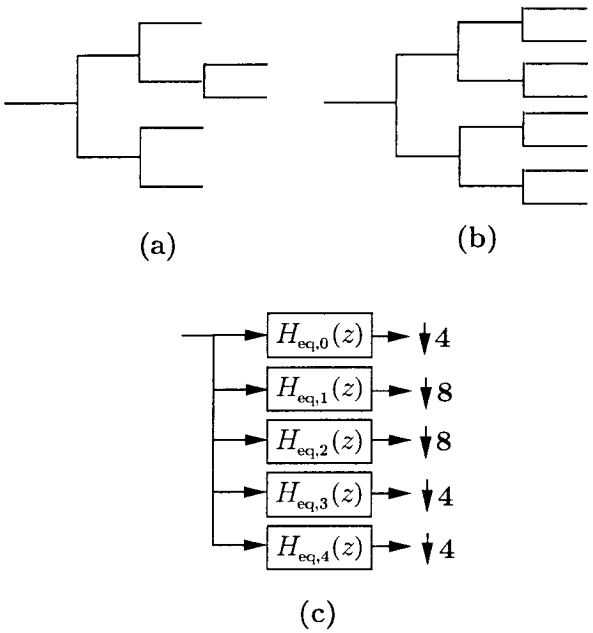


Fig. 2.8: (a) A wavelet packet decomposition, (b) the corresponding tree-structured filter bank, (c) nonuniform filter bank equivalent to (a).

used plain lines in this figure to represent branching of the tree structure. Each line represents a filter and downsampling by two as in Fig. 2.7. Wavelet packets can be considered as pruned versions of full tree-structured filter banks. The full tree-structured filter bank corresponding to the wavelet packet in Fig. 2.8a is shown in Fig. 2.8b. Also shown in Fig. 2.8c is a 5−channel nonuniform filter bank equivalent to the wavelet packet in Fig. 2.8a. By fixing the filters $H_0(z)$ and $H_1(z)$, there have been important developments to numerically design the optimal pruning of full tree-structured filter banks in the rate distortion sense [RVH96, Wic94]. Numerical optimization of filters $H_0(z)$ and $H_1(z)$ in such schemes has only recently been considered in [PMR97]. Theoretical optimization of filters together with such structures does not seem to be feasible.

One step to relax the optimization problem is to use different filter pairs for different

branches. Another step might be to use different numbers of channels (other than two) at different branches. In the extreme case, one can consider a nonuniform filter bank without any structure. Obviously, by doing this, one can obtain a nonuniform coder that achieves a better objective than wavelet or wavelet-packet based coders because the latter form special subclasses of the former. The price to pay is the lack of structure and therefore less efficient implementation of signal transformation. In this paper, we consider the theoretical optimization of orthonormal nonuniform filter banks for subband coding. Hence we impose the restriction that the filter bank is orthonormal, but we do not impose any constraints on the filters. In theory, biorthogonal systems can achieve better coding performance than orthonormal systems as in the special uniform case [VK98a]. However, we believe that solving the orthonormal case is a necessary first step for solving the biorthogonal case. In practice, the filters in wavelet and wavelet-packet based coding are designed to be biorthogonal, but they are made close to being orthonormal by careful scaling. Consider Fig. 2.9, where we show a subband

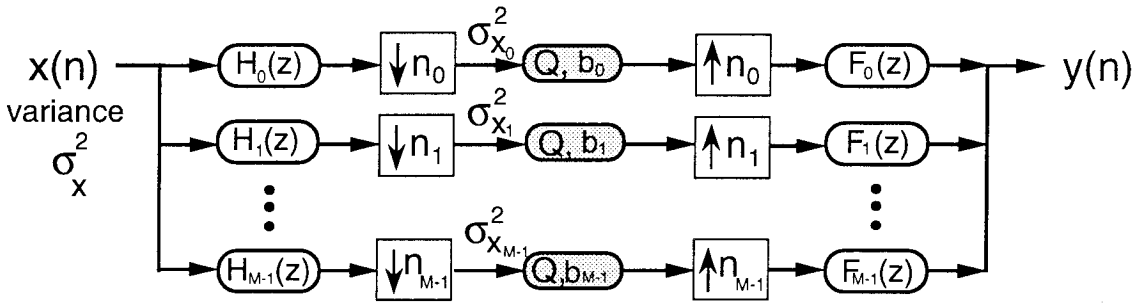


Fig. 2.9: Subband coding scheme with a nonuniform filter bank.

coding scheme with a nonuniform filter bank as its transform part. Given a number of channels M, we can have different decimation ratios n_j for different channels. If

$$\sum_{j=0}^{M-1} \frac{1}{n_j} = 1 \tag{2.25}$$

this corresponds to a maximally decimated system. Such a set will be called an admissible set. For every admissible set of decimation ratios, one can construct an orthonormal filter bank with perfect reconstruction property if we allow the filters to be ideal. If

we restrict ourselves to realizable filters, this is not the case and it is an open problem to find the class of admissible sets of decimation ratios that lead to realizable filters [DV94]. We will not be concerned about this issue in this chapter.

Let σ_x^2 be the original signal variance, $\sigma_{x_j}^2$ be the subband variances and $\sigma_{q_j}^2$ be the quantization noise variances in the subbands in Fig. 2.9. As in the uniform case, if we assume that the quantizers operate at high bit rates, then one can model them by a simple relationship: $\sigma_{q_j}^2 = c2^{-2b_j}\sigma_{x_j}^2$, where b_j is the number of bits at which the quantizer in jth channel is operating. If the total bit budget is b bits per sample, then we have

$$b = \sum_{j=0}^{M-1} \frac{b_j}{n_j} \tag{2.26}$$

Similar to the uniform case, by optimal bit allocation we have the following expression for coding gain [SV93]:

$$G_{coding} = \frac{\sigma_x^2}{\prod_{j=0}^{M-1}(\sigma_{x_j}^2)^{1/n_j}} \tag{2.27}$$

We have $\sigma_x^2 = \sum_{j=0}^{M-1} \sigma_{x_j}^2/n_j$ by the orthonormality of the filter bank. Hence the coding gain in the high bit rate case is the ratio of generalized arithmetic and geometric means of subband variances. In the uniform case where $n_j = M$, for all j, this reduces to the ratio of conventional arithmetic and geometric means and it is maximized by the Karhunen Loeve transform in the block transform case [HS63] and by optimal orthonormal filter banks as constructed in [Vai98] in the subband coding case with ideal filters.

2.3.1 Formulation for Arbitrary Bit Rates

If we do not assume that the quantizers operate at high bit rates, we can still formulate the nonuniform orthonormal subband coding problem as follows: consider Fig. 2.9 again. Let $\sigma_x^2, \sigma_{x_j}^2$, and $\sigma_{q_j}^2$ be variances of the original input, subband signals, and the quantization noise signals as defined before. Let $\mathcal{E}$ denote the reconstruction error of

this subband coding scheme. Then, by the orthonormality of the transform, we have:

$$\mathcal{E} = \sum_{j=0}^{M-1} \frac{\sigma_{q_j}^2}{n_j} \tag{2.28}$$

A filter bank is said to be optimum for subband coding if the reconstruction error $\mathcal{E}$ is minimized for a given total bit budget b as in (2.26). Only at high bit rates, optimality is the same as maximization of the coding gain expression in (2.27).

A General Quantizer Model

As in Sec. 2.1.2, we will model all the quantizers in Fig. 2.9 by a single quantization function $f(.)$ such that

$$\sigma_{q_j}^2 = f(b_j)\sigma_{x_j}^2 \tag{2.29}$$

So, as in the uniform case, **we do not assume high bit rates** but we do assume that all the channels have the same quantizer function $f(.)$. See the comments in Sec. 2.1.2. With this model, the reconstruction error of the subband coding scheme is

$$\mathcal{E} = \sum_{j=0}^{M-1} f(b_j)\frac{\sigma_{x_j}^2}{n_j} \tag{2.30}$$

2.3.2 Nonuniform Principal Component Filter Banks

Consider the problem of representation of signals by a subset of an $M-$channel nonuniform orthonormal filter bank: in Fig. 2.10, if we keep P of the subband signals without quantizing, and drop the other subbands, what is the filter bank that minimizes the reconstruction error? By the orthonormality of the transform, we can write

$$\mathcal{E} = \sum_{j=P}^{M} \frac{\sigma_{x_j}^2}{n_j} \tag{2.31}$$

Again by the orthonormality, this is minimized if $\sum_{j=0}^{P-1} \sigma_{x_j}^2/n_j$ is maximized. A filter bank that maximizes the partial sum $\sum_{j=0}^{P-1} \sigma_{x_j}^2/n_j$ for each P is called **a nonuniform principal component filter bank**. This is directly analogous to the definition of

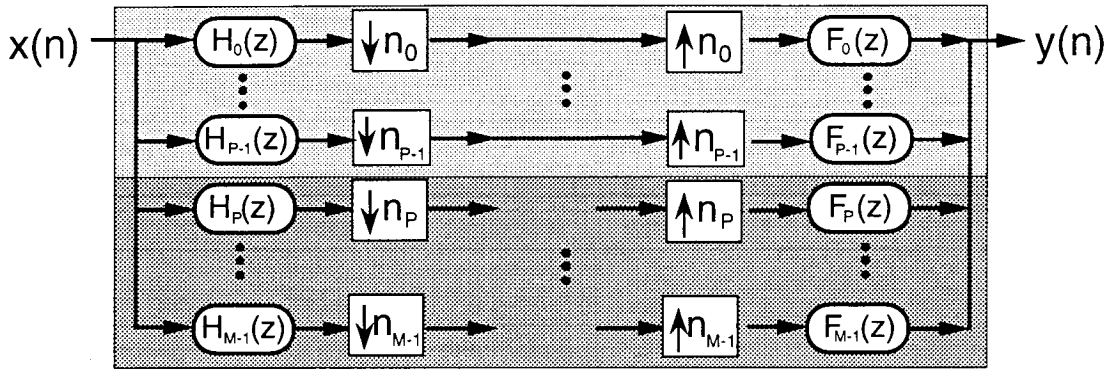


Fig. 2.10: Pertaining to the discussion of nonuniform principal component filter banks.

principal component filter banks in the uniform case. Notice, however, that a different ordering of the same set of decimation ratios in the nonuniform case results in a different PCFB. Hence the definition of a nonuniform PCFB should be made with respect to a particular ordering of an admissible set of decimation ratios:

Definition. A nonuniform orthonormal filter bank with a set of decimation ratios $\{n_j, j = 0, \ldots, M-1\}$ is said to be a PCFB for the permutation $\{j_i, i = 0, \ldots, M-1\}$; if among the class of nonuniform orthonormal filter banks that has the same set of decimation ratios, it maximizes the partial sum

$$\sum_{i=0}^{P-1} \frac{\sigma_{x_{j_i}}^2}{n_{j_i}} \tag{2.32}$$

for each $P = 1, \ldots, M$. Here $\sigma_{x_{j_i}}^2$ is the variance of the subband signal corresponding to the decimation ratio n_{j_i}.

2.3.3 Optimality Results for Nonuniform SBC

In the uniform case, we have shown in Sec. 2.1.4 that PCFB's are optimal orthonormal filter banks for all bit rates and bit allocation strategies. In the nonuniform case, the relationship between PCFB's and optimal orthonormal filter banks is not as strong due to the dependence of PCFB's on the ordering of decimation ratios. We start with the following observation:

Theorem 2. *Consider a nonuniform orthonormal SBC with a set of decimation ra-*

tios $\{n_j, j = 0, \ldots, M-1\}$ *and with a fixed corresponding bit allocation scheme:* $\{b_j, j = 0, \ldots, M-1\}$. *Let* $\{j_i, i = 0, \ldots, M-1\}$ *be the permutation such that* $f(b_{j_0}) \leq \ldots \leq f(b_{j_{M-1}})$. *Then a principal component filter bank for the permutation* $\{j_i, i = 0, \ldots, M-1\}$ *minimizes the total reconstruction error* $\mathcal{E}$ *of this SBC scheme.* $\diamondsuit$

Proof. We can write the total error as:

$$\begin{aligned} \mathcal{E} &= \sum_{j=0}^{M-1} f(b_j)\frac{\sigma_{x_j}^2}{n_j} = \sum_{i=0}^{M-1} f(b_{j_i})\frac{\sigma_{x_{j_i}}^2}{n_{j_i}} \\ &= \sum_{i=1}^{M-1} (f(b_{j_{i-1}}) - f(b_{j_i})) \sum_{l=0}^{i-1} \frac{\sigma_{x_{j_l}}^2}{n_{j_l}} + f(b_{j_{M-1}}) \sum_{l=0}^{M-1} \frac{\sigma_{x_{j_l}}^2}{n_{j_l}} \end{aligned} \tag{2.33}$$

The last term is equal to $f(b_{j_{M-1}})\sigma_x^2$ which is constant. The numbers $f(b_{j_{i-1}}) - f(b_{j_i})$ are nonpositive and fixed. Therefore, $\mathcal{E}$ is minimized if each of the partial sums $\sum_{l=0}^{i-1} \sigma_{x_{j_l}}^2 / n_{j_l}$ is maximized. A PCFB for the permutation $\{j_i, i = 0, \ldots, M-1\}$ achieves this by definition. ■

Comments. Such a PCFB is optimum for a subband coding scheme with a particular class of bit allocation strategies: namely those that satisfy $f(b_{j_0}) \leq \ldots \leq f(b_{j_{M-1}})$. In contrast to the uniform case, we do not know, in advance, which ordering corresponds to optimal bit allocation. However, we have the following result:

Theorem 3. *The optimal nonuniform orthonormal filter bank for subband coding with a given admissible set of decimation ratios* $\{n_j, j = 0, \ldots, M-1\}$ *is one of at most* $M!$ *PCFB's corresponding to all permutations of* $\{0, \ldots, M-1\}$. $\diamondsuit$

Proof. Let $b_j^*, j = 0, \ldots, M-1$ be the optimal bit allocation. Without loss of generality, assume that $f(b_{j_0}^*) \leq \ldots \leq f(b_{j_{M-1}}^*)$ for some permutation $\{j_i, i = 0, \ldots, M-1\}$. By Theorem 2, a PCFB for the permutation $\{j_i, i = 0, \ldots, M-1\}$ minimizes $\mathcal{E}$ for all such bit allocations. Hence a PCFB for this permutation is optimum for subband

coding with optimal bit allocation. This completes the proof. ■

Comments. As we stated before, for a given average bit rate b, we do not know the permutation of the optimal bit allocation and therefore we do not know which of the $M!$ PCFB's is optimum. But, given a quantizer function $f(.)$, one can enumerate minimum reconstruction errors of PCFB's corresponding to all permutations and select the best one. The optimum permutation may depend on both the quantizer function and the total bit budget.

For $M = 2$, there is only one admissible set of decimation ratios with one permutation, namely $\{2, 2\}$. Hence, there is only one PCFB which is uniform and therefore as proven in Sec. 2.1.4, it is optimum for subband coding for all bit rates and bit allocation strategies. For $M = 3$, we have the following illustrative example.

Example 5. For an input with a power spectral density (psd) as shown in Fig. 2.11 we want to design a 3−channel optimum nonuniform orthonormal filter bank with decimation ratios $\{6, 3, 2\}$. There are $3! = 6$ different PCFB's, one corresponding to each permutation. Details of construction of nonuniform PCFB's can be found in [KV98d]. For example, Fig. 2.11 shows three PCFB's corresponding to the permutations $\{6, 3, 2\}, \{6, 2, 3\}$ and $\{3, 6, 2\}$ respectively. Assume high bit rate assumptions hold so that the reconstruction error after the optimal bit allocation is of the form:

$$c2^{-2b}(\sigma^2_{x_{j_1}})^{1/n_{j_1}}(\sigma^2_{x_{j_2}})^{1/n_{j_2}}(\sigma^2_{x_{j_3}})^{1/n_{j_3}} \tag{2.34}$$

Ignoring the constant factor $c2^{-2b}$ for simplicity, the reconstruction error for the permutation $\{6, 3, 2\}$ is $\mathcal{E}_{632} = (\frac{1+c}{2})^{1/3}d^{1/2}$. Applying the same procedure for all permutations, we obtain:

$$\begin{aligned}
\mathcal{E}_{632} &= (\frac{1+c}{2})^{1/3}d^{1/2}, \\
\mathcal{E}_{623} &= (\frac{1+c+d}{3})^{1/2}d^{1/3}, \\
\mathcal{E}_{362} &= (\frac{1+c}{2})^{1/3}d^{1/2} = \mathcal{E}_{632},
\end{aligned}$$

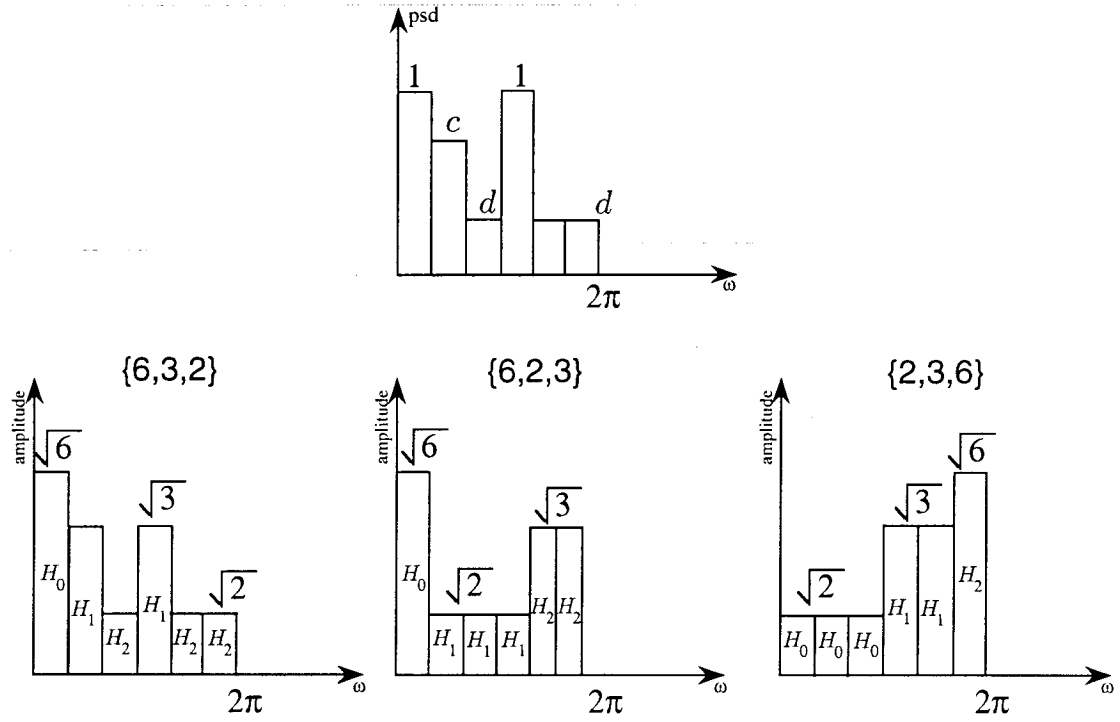


Fig. 2.11: Top: An input power spectral density. Bottom: PCFB's for the permutations $\{6,3,2\}, \{6,2,3\}$ and $\{3,6,2\}$ respectively.

$$\begin{aligned} \mathcal{E}_{326} &= (\frac{1+c}{2})^{1/3}(\frac{1+2d}{3})^{1/2}d^{1/6}, \\ \mathcal{E}_{263} &= (\frac{1+c+d}{3})^{1/2}d^{1/3} = \mathcal{E}_{623}, \\ \mathcal{E}_{236} &= (\frac{1+c+d}{3})^{1/2}(\frac{1+d}{2})^{1/3}d^{1/6} \end{aligned} \tag{2.35}$$

As an example, let $c = 1/2$ and $d = 1/4$. Then $\mathcal{E}_{632} = \mathcal{E}_{362}$ is the minimum. Hence a PCFB corresponding to permutation $\{6,3,2\}$ or $\{3,6,2\}$ is the optimum nonuniform orthonormal filter bank for this example.

Until this point, we have considered optimality of a nonuniform filter bank with a fixed set of decimation ratios. For a fixed number of channels M, one can expand the class of nonuniform orthonormal filter banks to that with all possible admissible sets of decimation ratios. An optimum filter bank among such a class will be referred to as **optimum $M-$channel nonuniform orthonormal filter bank**.

Corollary to Theorem 3. The optimum $M-$channel nonuniform orthonormal filter bank for subband coding is one of **finitely** many PCFB's corresponding to all admissible sets of M decimation ratios with all possible permutations.

Proof. We know that for a fixed set of decimation ratios, one PCFB corresponding to a particular permutation is optimum. Let us enumerate all admissible sets of decimation ratios and find the particular permutation for each set that minimizes the total reconstruction error. Clearly one of them has to have the minimum $\mathcal{E}$. The only thing that remains to be shown is the finiteness of the number of admissible sets of decimation ratios for a fixed number of channels M. This is done in Appendix A where we present an explicit recursive formula for the maximum possible decimation ratio m_M in all admissible sets for a given number of channels M. ■

Example 6. Let $M = 3$ and consider the input power spectrum shown in Fig. 2.12, where $\{a_i, i = 0, \ldots, 11\}$ is a decreasing set of positive numbers.

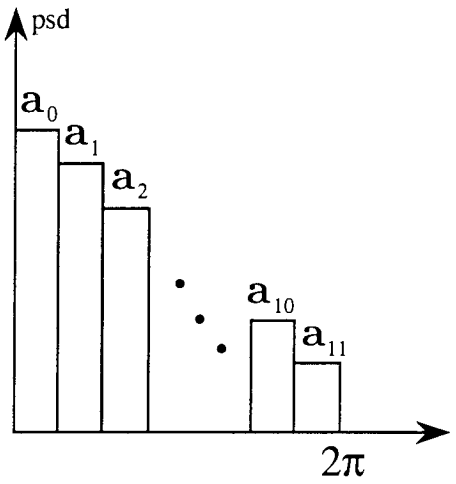


Fig. 2.12: Input power spectral density for Example 6.

From Appendix A, the maximum possible decimation ratio is $m_3 = 6$ and it is easy to verify that there are three admissible sets of decimation ratios, namely $\{3,3,3\}, \{4,4,2\}$, and $\{6,3,2\}$. For the set $\{3,3,3\}$ there is only one PCFB, while for $\{4,4,2\}$ there are 3 PCFB's, and for $\{6,3,2\}$ there are 6 PCFB's. Let $n_{j_0}, n_{j_1}, n_{j_2}$ be a particular permutation for a particular set of decimation ratios. Let the high bit rate

assumptions hold so that the reconstruction error is as in (2.34). Let $p_{j_i} = 12/n_{j_i}, i = 0, 1, 2$. Then ignoring the constant factor, the reconstruction error of the PCFB for the permutation $\{j_0, j_1, j_2\}$ is of the form:

$$\left[\left(\frac{1}{p_{j_0}} \sum_{i=0}^{p_{j_0}-1} a_i\right)^{p_{j_0}} \left(\frac{1}{p_{j_1}} \sum_{i=0}^{p_{j_1}-1} a_{p_{j_0}+i}\right)^{p_{j_1}} \left(\frac{1}{p_{j_2}} \sum_{i=0}^{p_{j_2}-1} a_{p_{j_0}+p_{j_1}+i}\right)^{p_{j_2}} \right]^{1/12}$$

Depending on the rate of decrease of numbers a_i, the optimum set of decimation ratios and the optimum permutation change. Let a and c be positive numbers such that $a_i > 0$ in the foregoing expressions. If $a_i = a - ci$, linearly decreasing, then we find that the PCFB for the ordered set $\{2, 3, 6\}$ is optimum. If $a_i = a\ (i + b)^{-c}$, polynomially decreasing, then we find that the PCFB for the ordered set $\{6, 3, 2\}$ is optimum. Finally, if $a_i = a^{-ci}$, exponentially decreasing, then we find that the uniform PCFB corresponding to $\{3, 3, 3\}$ is optimum. In [KV98d] we present results on the optimum permutation of decimation ratios depending on the rate of decrease of the input psd. On the practical side, most of the natural images have psd that have a polynomial type of decrease, and the wavelet-based coders use a tree-structured filter bank which is close to a PCFB for the permutation that is in decreasing order. For example, a 3-level wavelet-based coding typically uses a filter bank that is close to a PCFB for the ordered set $\{8, 8, 4, 2\}$.

APPENDIX

Proof that there are finitely many admissible sets for a given number of channels. An admissible set satisfies $\sum_{j=0}^{M-1} 1/n_j = 1$. Let m_M denote the maximum possible value that a decimation ratio in an admissible set can have. If we show that m_M is finite, then we are done. We show that this is the case by explicitly showing that

$$m_M = m_{M-1}(m_{M-1} + 1), \quad M = 2, 3, \ldots \tag{2.36}$$

with $m_1 = 1$. For $M = 1$, there is nothing to prove and for $M = 2$ we have only one set $\{2, 2\}$ and therefore $m_2 = 2$. For $M \geq 3$, the first $M - 1$ decimation ratios should be chosen to be as small as possible to maximize m_M. Similarly for $M - 1$, the first

$M-2$ decimation ratios should be as small as possible. Assume for $M-1$, we have found such a set, denote it by S_{M-1}. The largest element of S_{M-1} is m_{M-1}. For M, we must have $M-2$ elements of S_{M-1} excluding m_{M-1} and instead of m_{M-1} we should have next larger number which is $m_{M-1}+1$. Now we have the following identity:

$$\frac{m_{M-1}-1}{m_{M-1}}+\frac{1}{m_{M-1}+1}+\frac{1}{m_M}=1 \tag{2.37}$$

From this, the relation (2.36) follows. ■

Remark. From the above argument it also follows that for an arbitrary number of channels M, the set of decimation ratios that contain m_M is $S_M = \{m_1+1, m_2+1, \ldots, m_{M-1}+1, m_M\}$.

Chapter 3

Theory and Design of Optimum FIR Compaction Filters

3.1 Motivation

Consider the $M-$channel uniform orthonormal (or paraunitary) subband coder shown in Fig. 3.1. In terms of the filters we can express the orthonormality as [Vai93]

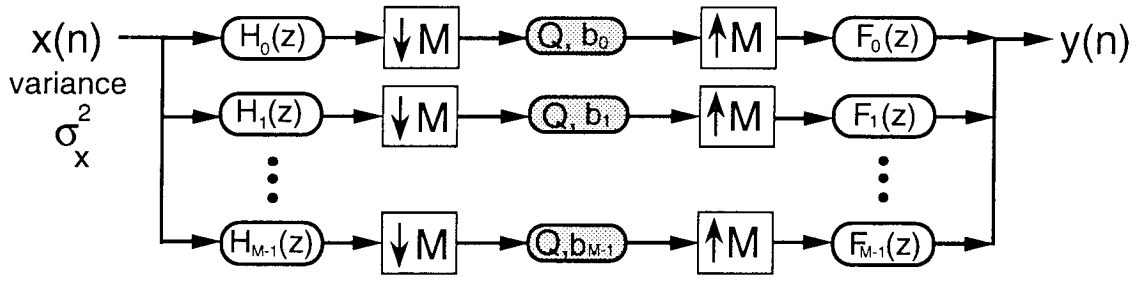


Fig. 3.1: $M-$channel uniform subband coder. Orthonormality implies $F_i(e^{j\omega}) = H_i^*(e^{j\omega})$, and $|H_i(e^{j\omega})|^2$ is Nyquist(M).

$$H_i(e^{j\omega})H_j^*(e^{j\omega})\Big|_{\downarrow M} = \delta(i-j) \tag{3.1}$$

This in particular implies that each filter satisfies the Nyquist(M) property (see Sec. 3.1.1 for notations and terminology):

$$|H_i(e^{j\omega})|^2\Big|_{\downarrow M} = 1 \tag{3.2}$$

If we assume that the quantizers operate at high bit rates, then from Sec. 2.1.1, with the optimal bit allocation, the coding gain is [JN84]:

$$G_{coding} = \frac{\sigma_x^2}{(\prod_{i=0}^{M-1} \sigma_{x_i}^2)^{1/M}} \tag{3.3}$$

where $\sigma_{x_i}^2$ is the variance at the output of $H_i(z)$, and $\sigma_x^2 = \sum_{i=0}^{M-1} \sigma_{x_i}^2/M$ by the orthonormality.

When some of the subband variances turn out to be smaller than a certain threshold, the corresponding channels should be dropped. In this case, the coding gain expression (3.3) is not applicable, and the total error is the sum of the quantization error and the error due to dropping. This is the case when high bit-rate assumption on the quantization noise sources is not satisfied.

In the high bit rate case, the optimum orthonormal filter bank that maximizes (3.3) is well-known for the case where filter orders are constrained to be less than M. This is the famous Karhunen-Loeve transform coder (KLT) and it diagonalizes the $M \times M$ **autocorrelation matrix** of the input. The solution for the case where the filter orders are unconstrained (ideal SBC) has also been mentioned in some detail in Chapter 2. The polyphase matrix [Vai93] of the solution diagonalizes the **psd matrix** in the frequency domain. This in particular implies the diagonalization of the autocorrelation matrix (which was both necessary and sufficient condition for the transform coding case). Diagonalization of the psd matrix at each frequency however, is not sufficient for the unconstrained filter bank to be optimum [Vai98]. There should be an additional ordering of the eigenvalues of the psd matrix at each frequency (spectral

majorization) [Vai98]. At a frequency ω, these eigenvalues are

$$\{S_{xx}(e^{j(\omega+k\frac{2\pi}{M})}),\ k=0,\ldots,M-1\} \tag{3.4}$$

where $S_{xx}(e^{j\omega})$ is the input psd.

For the arbitrary bit case, in Chapter 2, we have connected the problem of optimum uniform orthonormal subband coder to that of principal component representation of signals. At this point, the reader should review the results of Chapter 2 pertaining to the definition of PCFB's and their optimality for subband coding under arbitrary bit rates (not just high bit rates). If we think of the collection of the set of subband variances obtainable by *a certain class of orthonormal filter banks*, then the PCFB has the set of variances that majorizes every other set in the collection. The KLT and the optimal ideal orthonormal filter banks that maximizes the coding gain (3.3) are PCFB's in their corresponding classes. Therefore they are optimal for subband coding under arbitrary bit rates and bit allocation strategies (see Chapter 2).

A PCFB as defined in Chapter 2 maximizes the partial sums $\sum_{j=0}^{P-1}\sigma_{x_j}^2/M$ for each $P=1,\ldots,M$. In particular, the first filter $H_0(z)$ of a PCFB has the largest output variance compared to all the other filters that satisfy the Nyquist(M) property

$$|H_0(e^{j\omega})|^2\Big|_{\downarrow M}=1 \tag{3.5}$$

Such a filter is called an optimal compaction filter and it is the subject of this chapter to investigate its analysis and design. Of course, if there are any constraints on the class of filter banks, the first filter has to be such that it can be completed to an orthonormal filter bank in that class. For example, if the filter bank is order constrained, then the first filter has to be order constrained accordingly.

Although in the transform coding case and the ideal case a PCFB does exist, in the intermediate case (i.e., finite order filter banks), unfortunately the existence of such a filter bank is not always guaranteed. The reader is referred to Chapter 2 for examples where no FIR PCFB of a given order or degree exists for a particular input

psd. However, if it exists, then designing an optimum FIR compaction filter $H_0(e^{j\omega})$ is the first step of finding such a filter bank. In that case, Moulin et al. [MAKP96] uses a result due to Vaidyanathan el al. [VNDS89] to optimally complete the filter bank. This is based on the fact that, if one filter $H_0(e^{j\omega})$ in an FIR orthonormal filter bank of a given degree is known, then the number of freedoms available for the design of the remaining filters is limited. This remaining freedom can in fact be captured with a simple constant unitary matrix $\mathbf{U}$. Essentially the last $M-1$ rows of $\mathbf{U}$ are free and should be chosen to maximize the coding gain. The optimum $\mathbf{U}$ is the KLT corresponding to its input vector which is determined by the first filter $H_0(z)$ and the original input $x(n)$. As we will see, the optimality of a compaction filter depends only on its magnitude-squared frequency response. Hence, for an optimum magnitude-squared frequency response, one has the choice of selecting a particular spectral factor. It turns out that this choice affects the coding gain (see Chapter 2 for examples), and one has to choose the best spectral factor.

For the two-channel case, the existence of a PCFB is assured even if the filters are order-constrained. To see this, note that a two-channel PCFB maximizes only $\sigma_{x_0}^2$. By orthonormality, the sum $\sigma_{x_0}^2 + \sigma_{x_1}^2$ is constant. Once one order-constrained filter that maximizes $\sigma_{x_0}^2$ is found, all that remains is to find another filter such that the two filters form an orthonormal filter bank. It is very well known that the second filter is determined from the first filter by simple flipping and sign changes (see (3.39)). Hence in the two channel case, designing an optimal FIR compaction filter is the same as designing an optimal FIR orthonormal filter bank for subband coding. By the results of Chapter 2, the optimality of this filter bank is independent of the bit rates involved. In the high-bit rate case, the coding gain expression becomes

$$G_{coding} = \frac{\sigma_x^2}{\sqrt{\sigma_{x_0}^2 \sigma_{x_1}^2}} = \frac{\sigma_x^2}{\sqrt{\sigma_{x_0}^2 (2\sigma_x^2 - \sigma_{x_0}^2)}} \tag{3.6}$$

In this case we can write

$$G_{coding} = \frac{1}{\sqrt{G_{comp}(2 - G_{comp})}} \tag{3.7}$$

where $G_{comp} = \sigma_{x_0}^2/\sigma_x^2$ is the compaction gain defined more precisely in Sec. 3.2.

In this chapter, we focus on the design of an optimum FIR compaction filter when the order is such that $M < N < \infty$. As we discussed, for $M = 2$, this is equivalent to the design of optimum orthonormal filter banks for subband coding, and with trivial extensions, to the design of optimal wavelet generating filters. For arbitrary M, the design in [MM98] can be used to obtain a good orthonormal filter bank using the compaction filter. The usefulness of signal-adapted designs in image coding with the mean-squared error as the criterion is demonstrated in [DMS92, TZ92, TDK95].

Other Applications of Compaction Filters

In view of principal component analysis, in addition to subband coding and data compression, other immediate applications of compaction filters are signal modeling and model reduction, low-resolution data representation (multimedia databases), and classification. Two other interesting applications of compaction filters are adaptive echo cancellation [JLW96] and time-varying system identification [TG93]. Consider the design of zero intersymbol-interference (ISI) transmitter and receiver filters for data transmission over bandlimited communication channels. Let $H(e^{j\omega})$ and $F(e^{j\omega})$ be the transmitter and receiver filters respectively. To maximize SNR in the presence of additive white noise, **matched filters** should be used, that is $F(e^{j\omega}) = H^*(e^{j\omega})$. With this, the zero ISI property becomes $|H(e^{j\omega})|^2\big|_{\downarrow M} = 1$ which is nothing but Nyquist(M) property! Such optimally designed filters are used, for example, in voiceband data modem applications [CU82].

3.1.1 Notations and Terminology

1. The notations $\tilde{X}(z)$, $X(z)|_{\downarrow M}$, and the Nyquist(M) property are defined in Sec. 1.4.

2. The notation $x_L(n)$ stands for a periodic sequence with periodicity L. If there is a reference to a finite sequence $x(n)$ as well, then it is to be understood that $x_L(n)$ is the periodical expansion of $x(n)$, i.e., $x_L(n) = \sum_{i=-\infty}^{\infty} x(n+Li)$. The Fourier series coefficients (FSC) of $x_L(n)$ is denoted by $X_L(k)$. For L a multiple of M, a periodic sequence $x_L(n)$ is said to be Nyquist(M) if

$$x_L(Mn) = \delta_K(n) \triangleq \sum_{i=-\infty}^{\infty} \delta(n+Ki) \tag{3.8}$$

where $K = L/M$.

3. Positive definite sequences. Let a sequence $\{x(n), n = 0, \ldots, N\}$ be given and let $\mathbf{P}$ be the Hermitian Toeplitz matrix whose first row is $[x(0)\ x(1)\ \ldots\ x(N)]$. The sequence $\{x(n)\}$ is called positive definite if $\mathbf{P}$ is positive definite. Let $[a(0)\ a(1)\ \ldots\ a(N)]^T$ denote an eigenvector corresponding to the maximum eigenvalue of $\mathbf{P}$. Then the filter $A(z) = \sum_{n=0}^{N} a(n)z^{-n}$ is called a **maximal eigenfilter** of $\mathbf{P}$. The definitions for negative definite sequences and minimal eigenfilters are analogous.

3.1.2 New Results and Outline of the Chapter

In Sec. 3.2, we formulate the optimum FIR energy compaction problem and present a brief review of existing work. The remaining sections contain new results: In Sec. 3.3, we give an extension of the technique in [ADM95] for the analytical solution of the FIR energy compaction problem in the two-channel case. This is equivalent to the problem of optimal two-channel orthonormal filter bank that maximizes the coding gain and with a trivial extension (constraining some zeros at $\omega = \pi$), to the optimum wavelet generating filter problem. The method involves Levinson recursion and two spectral factorizations of half the filter order. We will see that the **analytical method** is related to the well-known line-spectral theory in signal processing society [RM87].

We develop a new technique called the **window method** for the design of FIR compaction filters for the $M-$channel case (Sec. 3.4). The window method has the advantage that no optimization tools or iterative numerical techniques are necessary. The solution is generated in a finite number of elementary steps, the crucial step being a simple comparison operation on a **finite frequency grid**.

We discuss some drawbacks of the LP method and propose some improvements (Sec. 3.4). We also consider the design and implementation of compaction filters in multiple stages (Sec. 3.5). Similar to the case of IFIR filters in filter design practice [NCYM84, Vai93], this is very efficient both in the design stage and in implementation. *MATLAB programs can be found at our webpage* [htt] *for the algorithms described in this chapter.*

The three techniques (the analytical method, the window method, and the LP method) are complementary rather than competing with each other. For the two-channel case, the analytical method should be the choice whenever it is successful. If it is not or if $M > 2$, for high filter orders the window method should be preferred. If the filter orders are low, then linear programming should also be considered, though sometimes the window method performs as good as LP even for low filter orders (see Example 12).

The results of this chapter have been published in a special issue of a journal [KV98e] and at various conferences [KV97a, KV97b, KV96a].

3.2 The FIR Energy Compaction Problem

An FIR filter $H(z)$ of order N will be called a valid **compaction filter** for the pair (M, N) if $|H(e^{j\omega})|^2$ is Nyquist(M) that is, $|H(e^{j\omega})|^2\big|_{\downarrow M} = 1$. Let $G(e^{j\omega}) = |H(e^{j\omega})|^2$. We will call $G(z)$ the **product filter** corresponding to $H(z)$. Conversely, $G(z)$ is the product filter of a valid compaction filter for the pair (M, N) if it is of **symmetric order** N, that is $G(z) = \sum_{n=-N}^{N} g(n)z^{-n}$ and satisfies the following two conditions:

$$g(Mn) = \delta(n) \text{ (Nyquist}(M) \text{ condition)} \quad \text{and} \quad G(e^{j\omega}) \geq 0 \text{ (nonnegativity)} \tag{3.9}$$

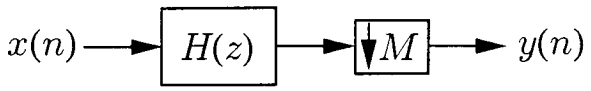


Fig. 3.2: $M-$channel compaction filter. $|H(e^{j\omega})|^2$ is Nyquist(M).

Now consider Fig. 3.2 where $H(z)$ is applied to a zero-mean WSS input $x(n)$ with psd $S_{xx}(e^{j\omega})$, and the output is decimated by M. The optimum FIR compaction problem is to find a valid compaction filter $H(z)$ for the pair (M, N) such that the variance σ_y^2 of $y(n)$ is maximized. Since decimation of a WSS process does not alter its variance, we have

$$\sigma_y^2 = \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} G(e^{j\omega}) S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi} \tag{3.10}$$

We define the **compaction gain** as

$$G_{comp}(M, N) = \frac{\sigma_y^2}{\sigma_x^2} = \frac{\int_{-\pi}^{\pi} G(e^{j\omega}) S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi}}{\int_{-\pi}^{\pi} S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi}} \tag{3.11}$$

The aim therefore is to maximize the compaction gain under the constraints (3.9).

Two Extreme Special Cases

Let us consider the two special cases: (a) the case where $N < M$ and (b) the ideal case, $N = \infty$. In the first case, the condition $g(Mn) = \delta(n)$ is the same as $g(0) = 1$. This is equivalent to saying that $H(e^{j\omega})$ has unit energy. Let $\mathbf{h}$ be $(N+1) \times 1$ vector formed by $h(n)$ and let $\mathbf{R}_{xx}$ be the $(N+1) \times (N+1)$ autocorrelation matrix of $x(n)$. Then the problem is to maximize $\mathbf{h}^\dagger \mathbf{R}_{xx} \mathbf{h}$ subject to the condition $\mathbf{h}^\dagger \mathbf{h} = 1$. By Rayleigh's principle [HJ85], the optimum $\mathbf{h}$ is the **maximal eigenvector** of $\mathbf{R}_{xx}$. In other words, $H(z)$ is the maximal eigenfilter of $\mathbf{R}_{xx}$. Let the maximum eigenvalue of $\mathbf{R}_{xx}$ be denoted by $\lambda_{max}\{r(n)\}_0^N$, where $r(n)$ is the input autocorrelation sequence. Then the optimum compaction gain is $\lambda_{max}\{r(n)\}_0^N / \sigma_x^2$. The second case has the following solution [TG95, Uns93a, Vai96]: if we write $H(z)$ in polyphase form [Vai93], $H(z) = \sum_{k=0}^{M-1} z^{-k} E_k(z^M)$ and if $\mathbf{S}_{xx}(e^{j\omega})$ denotes the $M \times M$ psd matrix of $x(n)$, then

for each ω,

$$\mathbf{e}(e^{j\omega}) = [E_0(e^{j\omega}) \dots E_{M-1}(e^{j\omega})]^\dagger \tag{3.12}$$

is the **maximal eigenvector** of $\mathbf{S}_{xx}(e^{j\omega})$. Equivalently for each $\omega \in [0, \frac{2\pi}{M})$, let $S_{xx}(e^{j(\omega+i_0\frac{2\pi}{M})})$ be the maximum of the set

$$\{S_{xx}(e^{j(\omega+i\frac{2\pi}{M})}),\ i = 0, 1, \dots, M-1\} \tag{3.13}$$

Then $H(e^{j(\omega+i_0\frac{2\pi}{M})}) = \sqrt{M}$ and $H(e^{j(\omega+i\frac{2\pi}{M})}) = 0$ for $i \neq i_0$. Note that the eigenvalues of $\mathbf{S}_{xx}(e^{j\omega})$ are $\{S_{xx}(e^{j(\omega+i\frac{2\pi}{M})}),\ i = 0, 1, \dots, M-1\}$. Let $G_{ideal}(M)$ denote the corresponding compaction gain. We can write

$$G_{ideal}(M) = M \int_\Omega S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi} / \sigma_x^2 \tag{3.14}$$

where Ω is the passband of $H(e^{j\omega})$.

In the Nth order FIR compaction problem, we do not have the flexibility of assigning values to $H(e^{j\omega})$ independently for each ω. This is because $H(e^{j\omega})$ is determined by its $N+1$ frequency samples. For $N > M$, the problem is not an eigenfilter problem either, as the condition $g(Mn) = \delta(n)$ implies more than the simple unit-energy condition. In Sec. 3.4 we will introduce a suboptimal method called the window method. Interestingly enough, the method involves two stages that can be associated with the above special cases. While the method is suboptimal, it produces compaction gains very close to the optimum ones especially for high filter orders.

Upper Bounds on the Compaction Gain

We have the following bounds for the compaction gain:

$$G_{opt}(M,N) \leq \lambda_{max}\Big\{r(n)\Big\}_0^N, \quad G_{opt}(M,N) \leq G_{ideal}(M), \quad \text{and} \quad G_{opt}(M,N) \leq M \tag{3.15}$$

For the first inequality, let k be an integer such that $kM > N$. From the first special case above we have $G_{opt}(kM,N) = \lambda_{max}\Big\{r(n)\Big\}_0^N$. Since Nyquist($M$) property implies

Nyquist(kM) property, $G_{opt}(M,N) \leq G_{opt}(kM,N)$. The second inequality follows because $G_{opt}(M,N) \leq G_{opt}(M,N+1)$. For the last inequality, first observe that $G(e^{j\omega}) \leq M$. Hence,

$$\sigma_y^2 = \int_{-\pi}^{\pi} G(e^{j\omega}) S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi} \leq M \int_{-\pi}^{\pi} S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi} = M\sigma_x^2 \tag{3.16}$$

The equality holds if and only if $G(e^{j\omega}) = M$ for all ω for which $S_{xx}(e^{j\omega}) \neq 0$. If $S_{xx}(e^{j\omega})$ is not line-spectral, this requires $G(e^{j\omega})$ to be identically zero for some region of frequency which is impossible since the order is assumed to be finite. Hence for a process that is not line-spectral, the last inequality is strict. That is, $G_{opt}(M,N) < M$. For $M = 2$, we will derive another upper bound for $G_{opt}(2,N)$ in Sec. 3.3.2 (see (3.31)) for a class of random processes. Whenever the analytical method of Sec. 3.3 succeeds, this bound is in fact achieved.

3.2.1 Previous Work

Here is a brief review of the existing methods for FIR compaction filter design:

1. **Lattice Parameterization.** Two-channel real-coefficient orthonormal filter banks can be completely parameterized by a lattice structure [Vai93]. Each stage in the lattice has an angle parameter θ_k. The objective function, however, is highly nonlinear function of these angles. Delsarte et al. [DMS92] propose an iterative algorithm called the ring algorithm to optimize the lattice stage by stage. Taubman and Zakhor [TZ92] propose an algorithm aimed at finding a globally optimum solution for small filter orders. They extend the results to two-dimensional nonseparable filters as well.

2. **Quadratically Constrained Optimization.** One can formulate the problem in terms of the compaction filter impulse response $h(n)$: maximize $\mathbf{h}^\dagger \mathbf{R}_{xx} \mathbf{h}$ subject to $\mathbf{h}^\dagger \mathbf{A}^i \mathbf{h} = \delta(i), \quad i = 0, \ldots, K$, where $\mathbf{A}$ is an appropriately chosen singular matrix, and $\mathbf{A}^0 = \mathbf{I}$. Here $\mathbf{R}_{xx}$ is the input autocorrelation matrix, and $\mathbf{h}$ is the vector of filter coefficients. The authors in [CLA91, Van92] use Lagrangian

techniques to solve the problem for the two-channel case and for small filter orders. Chevillat and Ungerboeck [CU82] provide an iterative projected gradient algorithm for the $M-$channel case and for moderate filter orders.

3. **Eigenfilter Method.** In [VNDS89], the authors design one filter of an $M-$channel orthonormal filter bank using the so-called eigenfilter method. The objective in their design is to have a good frequency response. However, one can modify the technique to incorporate the input statistics. This can be done by using the psd $S_{xx}(e^{j\omega})$ as a weighting function in the optimization. The paper also discusses how to design a good orthonormal filter bank using the remaining degrees of freedom. In [MAKP96] Moulin et al. show how to use this idea for the statistical optimization of orthonormal filter banks.

4. **Linear programming.** The objective is a linear function of the impulse response $g(n)$ of $G(e^{j\omega}) = |H(e^{j\omega})|^2$. The Nyquist($M$) property can be trivially achieved. However, we need to impose $G(e^{j\omega}) \geq 0$ for all ω. This can be written as a linear inequality for each ω in terms of $g(n)$. Hence the problem is a linear programming (LP) problem with infinitely many inequality constraints, hence the name semi-infinite programming (SIP). Although we used LP independently to design compaction filters at the early stages of this project, it was first proposed and examined in depth by Moulin et al. [Mou95, MAKP96, MAKP97].

5. **Analytical methods.** Aas et al. [ADM95] worked on a closely related problem for the two-channel case. They have constructed a Nyquist(2) real filter $H(e^{j\omega})$ that maximizes the baseband energy $\int_{-\pi/2}^{\pi/2} |H(e^{j\omega})|^2 \frac{d\omega}{2\pi}$. Based on the fundamentals of Gaussian quadrature, the authors were able to obtain an analytical method to identify the unit-circle zeros of $H(z)$ which uniquely determine it. In this chapter, this method will be referred to as **analytical method**. In Sec. 3.3 we present extensions of the analytical method. While the original method primarily addresses conventional half-band filter design, we will show how to adapt the idea for the case of FIR compaction filter design for a given input psd. Interestingly enough, we shall show that the analytical method is related to the well-known

line-spectral theory in signal processing society [RM87]. An analytical expression for the compaction gain for $N = 3$ is presented in [UO95]. See also [SdS96] for $N = 3$.

The major disadvantage of the first three methods is that they are iterative and there is a possibility of reaching a locally optimum solution. Nonlinearity of the objective is very severe in the first technique. A milestone in the design approaches is the formulation of the problem in terms of the product filter. This is done in the last two methods above. In this chapter, we also design the product filters. A spectral factorization step is necessary to find the compaction filter coefficients in contrast to the first three methods. In a newly developed technique, Tuqan and Vaidyanathan [TV98] uses state space theory to cast the problem into a semi definite programming problem. The formulation is such that the spectral factorization is automatically achieved within the algorithm.

3.3 Analytical Method

In this section we consider the special case of two channels ($M = 2$) and assume that the input $x(n)$ is real so that the compaction filter coefficients $h(n)$ can be assumed to be real. For this two-channel case we will show that the optimal product filter $G(e^{j\omega})$ can sometimes be obtained using an analytical method instead of going through a numerical optimization procedure. We will also present a number of examples which demonstrate the usefulness of the method. Also presented are examples where the analytical method can be shown to fail. As in [ADM95], one can modify the algorithms of this section to constrain the filters to have specified number of zeros at $\omega = \pi$ to generate optimal wavelets.

The analytical method is motivated by the fact that, under some conditions to be explained, the objective function (3.10) can be conveniently expressed as a summation over a finite number of frequencies determined by the psd $S_{xx}(e^{j\omega})$. The summation involves the samples of a modified polyphase component of $G(e^{j\omega})$. This will allow us to optimize the modified polyphase component, and hence $G(z)$, essentially *by inspection.*

Using these observations, we come up with an algorithm that determines the unit-circle zeros of the compaction filter. Using the Nyquist(2) condition, this in turn determines the filter itself.

The inspiration for our work in this section comes from the recent contribution by Aas et al. [ADM95] where the Gaussian quadrature technique is cleverly used to address the problem of maximizing the baseband energy of half-band filters. Our work in this section differs in a number of respects. First we do not use Gaussian quadrature, but take advantage of an elegant representation for positive definite sequences which results from the theory of line-spectral processes. Second, we take into account the knowledge of the input psd in the optimization process. We give the analytical solutions for some practically important classes of random processes.

Let us represent the product filter $G(z) = \sum_{n=-N}^{N} g(n)z^{-n}$ in the traditional polyphase form [Vai93] for $M = 2$: $G(z) = E_0(z^2) + z^{-1}E_1(z^2)$. By the Nyquist(2) property we have $E_0(z) = 1$. For the real coefficient case we have $g(n) = g(-n)$, and it follows that the coefficients of the FIR filter $E_1(z)$ have the symmetry demonstrated in Fig. 3.3.

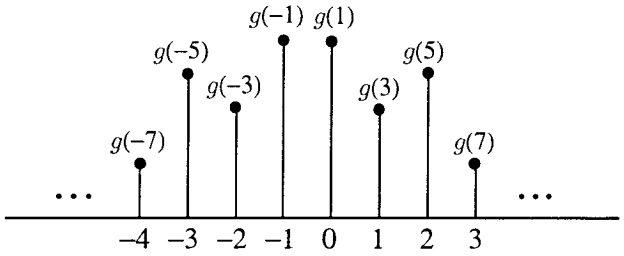


Fig. 3.3: Coefficients of the polyphase component $E_1(z)$ of the product filter $G(z)$. Because of the symmetry $g(n) = g(-n)$, we have $E_1(-1) = 0$.

This implies, in particular, that $E_1(z) = 0$ for $z = -1$. By factoring the zero at $z = -1$, we can write $E_1(z) = (1+z)G_1(z)$ where $G_1(z)$ has symmetric real coefficients. Hence we can write

$$G(z) = 1 + (z + z^{-1})G_1(z^2), \quad \text{i.e.,} \quad G(e^{j\omega}) = 1 + 2\cos\omega G_1(e^{j2\omega}) \tag{3.17}$$

Since Nyquist condition and nonnegativity of $G(e^{j\omega})$ together imply $0 \leq G(e^{j\omega}) \leq 2$,

the modified polyphase component $G_1(e^{j\omega})$ is bounded as follows:

$$-\frac{1}{2\cos(\omega/2)} \leq G_1(e^{j\omega}) \leq \frac{1}{2\cos(\omega/2)}, \quad -\pi < \omega < \pi \tag{3.18}$$

Notice that $G(z)$ and $G_1(z)$ can be determined from each other uniquely. We shall express the output variance σ_y^2 in terms of $G_1(e^{j\omega})$ so that we can see how to optimize the coefficients of $G_1(z)$. For this, write the input psd in the traditional polyphase form as $S_{xx}(z) = S_0(z^2) + z^{-1}S_1(z^2)$. Then σ_y^2 can be simplified into the form

$$\sigma_y^2 = r(0) + \int_{-\pi}^{\pi} G_1(e^{j\omega})\Psi_x(e^{j\omega})\frac{d\omega}{2\pi} \tag{3.19}$$

where $\Psi_x(z) = (1 + z^{-1})S_1(z)$ or equivalently

$$\Psi_x(e^{j\omega}) = \cos(\omega/2)\Big(S_{xx}(e^{j\omega/2}) - S_{xx}(e^{j(\pi-\omega/2)})\Big) \tag{3.20}$$

Using Parseval's relation the objective can be written as

$$\sigma_y^2 = r(0) + \sum_{n=-(N-1)/2}^{(N-1)/2} g_1(n)\psi_x(n) \tag{3.21}$$

where $\psi_x(n)$ is the inverse transform of $\Psi_x(z)$ which is produced below explicitly for convenience.

$$\psi_x(0) = 2r(1),\ \psi_x(1) = r(1) + r(3),\ \ldots,\ \psi_x(\frac{N-1}{2}) = r(N-2) + r(N), \tag{3.22}$$

and $\psi_x(n) = \psi_x(-n)$. Our aim is to maximize the second term in (3.21) for fixed $\psi_x(n)$ (i.e., fixed input) by choosing $g_1(n)$ under the constraint (3.18) and the usual filter-order constraint. Under the assumption that the input-dependent sequence $\psi_x(n)$ is **positive or negative definite** (see Sec. 3.1.1 for definition), we will show how this can be done analytically. The significance of this assumption on $\psi_x(n)$ is explained in Sec. 3.3.3. We will need the representation theorem of the next section for positive definite sequences.

3.3.1 Representation of Positive Definite Sequences

Theorem 1. *Given a positive definite sequence of $m+1$ complex numbers $\{\phi(n),\ n = 0,\ldots,m\}$, there exists a representation of the form*

$$\phi(n) = \sum_{k=0}^{m} \alpha_k e^{j\omega_k n}, \quad n = 0,\ldots,m \tag{3.23}$$

where $\alpha_k > 0,\ k = 0,\ldots,m$, and ω_k's are all distinct. $\diamondsuit$

Comments. Note that this is different from the Caratheodory representation theorem which is the basis for the Pisarenko method [Pis73] for identifying sinusoidal signals under noise: given $\{\phi_n,\ n = 1,\ldots,m\}$ there exists a representation of the form $\phi_n = \sum_{k=1}^{m} \alpha_n e^{j\omega_k n},\ n = 1,\ldots,m$, where α_n's are nonnegative. The frequencies ω_n's are the angles of the unit magnitude roots of the minimal eigenpolynomial of a matrix $\mathbf{Q}$. The matrix $\mathbf{Q}$ is $(m+1)\times(m+1)$ Hermitian Toeplitz with the first row $[\phi_0\ \phi_1\ldots\phi_m]$ where $\phi_0 = \sum_{k=1}^{m} \alpha_k$ is the positive number that makes the matrix singular. Here, the number of distinct frequencies depends on the multiplicity of the minimum eigenvalue of the so obtained matrix. If the multiplicity is 1, there are m distinct frequencies. If we start with a positive definite sequence $\phi_n,\ n = 0,\ldots m$, then Caratheodory representation takes the form:

$$\phi_n = (\phi_0 - \sum_{k=1}^{m} \alpha_k)\delta(n) + \sum_{k=1}^{m} \alpha_k e^{j\omega_k n}, \quad n = 0,\ldots,m \tag{3.24}$$

This is obviously not the same as (3.23) and is not suitable for our purposes. Although Theorem 1 turns out to be well-known in the literature [AK62], we include our own proof below for two reasons: i) it is elegant and uses the theory of line-spectral processes, and ii) it reveals us the algorithmic steps of the analytical method.

Proof of Theorem 1. Let $\mathbf{P}$ be the $(m+1)\times(m+1)$ Hermitian Toeplitz matrix whose first row is $\Phi^T = [\phi(0)\ \phi(1)\ldots\phi(m)]$. Consider the extension of $\mathbf{P}$ into a singular $(m+2)\times(m+2)$ Hermitian Toeplitz matrix $\hat{\mathbf{P}}$ such that its $(m+1)\times(m+1)$ principal submatrix is $\mathbf{P}$. This extension is merely augmenting an extra element

$\phi(m+1)$ to the end of Φ and forming the corresponding Hermitian Toeplitz matrix. The number $\phi(m+1)$ is chosen to make $\hat{\mathbf{P}}$ singular. This can always be done because of the following reason: for the positive definite matrix $\mathbf{P}$, one can run the well-known Levinson recursion procedure [RM87] to obtain the optimal mth order predictor polynomial $A_m(z)$. If one now considers the following continuation of the recursion $P_c(z) = A_m(z) + cz^{-(m+1)}\tilde{A}_m(z)$ with $|c| = 1$, then this corresponds to the singular predictor polynomial of a random process with singular autocorrelation matrix $\hat{\mathbf{P}}$. The result now follows from a well established fact [RM87] which states that a WSS process is line spectral with exactly $m+1$ lines if and only if its $(m+1) \times (m+1)$ autocorrelation matrix is nonsingular and $(m+2) \times (m+2)$ autocorrelation matrix is singular. Applying this result to a process with autocorrelation matrix $\hat{\mathbf{P}}$, we get (3.23). ■

Remarks. It is clear that $P_c(z)$ defined in the above proof is also the minimal eigenfilter of $\hat{\mathbf{P}}$. The zeros of $P_c(z)$ are all on the unit circle and distinct. Let $\{e^{j\omega_k},\ k = 0, \ldots, m\}$ be these zeros. The distinct frequencies $\{\omega_k, k = 0, \ldots, m\}$ are referred to as the line-spectral frequencies and α_k is the power at the frequency ω_k. The representation (3.23) is not unique because of the nonuniqueness of the unit magnitude constant c in the proof.

Real Case. For real $x(n)$, the predictor polynomial $A_m(z)$ and the constant c are real. Hence we have two cases: $c = \pm 1$. The case $c = 1$ leads to a symmetric polynomial $P_1(z)$, while the case $c = -1$ leads to an antisymmetric polynomial $P_{-1}(z)$. It is a well-known fact that the distinct unit-circle zeros of these two polynomials are interleaved. For simplicity assume that m is odd. Then $P_{-1}(z)$ has both of the zeros $z = 1$ and $z = -1$ and $P_1(z)$ has none of them. Using $P_1(z)$, we have the following representation for a real positive definite sequence $\phi(n)$:

$$\phi(n) = \sum_{k=0}^{(m-1)/2} \beta_k \cos \omega_k n, \quad n = 0, \ldots, m \tag{3.25}$$

where $\beta_k > 0,\ k = 0, \ldots, (m-1)/2$, and ω_k's are all distinct and $0 < \omega_k < \pi,\ k = 0, \ldots, (m-1)/2$.

3.3.2 Derivation of the Analytical Method

Assume for simplicity $(N-1)/2$ is odd and assume $\{\psi_x(n),\ n=0,\ldots,(N-1)/2\}$ is positive definite. Applying the real form of the representation we have

$$\psi_x(n) = \sum_{k=0}^{(N-3)/4} \beta_k \cos\omega_k n, \quad n=0,\ldots,\frac{N-1}{2} \tag{3.26}$$

The objective (3.21) can therefore be written as

$$\sigma_y^2 = r(0) + \sum_{k=0}^{(N-3)/4} \beta_k \sum_{n=-(N-1)/2}^{(N-1)/2} g_1(n)\cos\omega_k n = r(0) + \sum_{k=0}^{(N-3)/4} \beta_k G_1(e^{j\omega_k}) \tag{3.27}$$

From (3.18), the output variance (3.27) is maximized if

$$G_1(e^{j\omega_k}) = \frac{1}{2\cos(\omega_k/2)}, \quad k=0,\ldots,(N-3)/4 \tag{3.28}$$

This implies $G(e^{j\omega_k/2}) = 2$, and by Nyquist(2) property

$$G(e^{j(\pi-\omega_k/2)}) = 0, \quad k=0,\ldots,(N-3)/4. \tag{3.29}$$

Notice that these zeros are all located in the region $(\pi/2,\pi)$. Since $0 \le G(e^{j\omega}) \le 2$, the derivatives of $G(e^{j\omega})$ should vanish at the above frequencies. Hence we should have $G'(e^{j\omega_k/2}) = 0,\ k=0,\ldots,(N-3)/4$. In view of (3.17), this in turn implies

$$G_1'(e^{j\omega_k}) = \frac{\sin(\omega_k/2)}{4\cos^2(\omega_k/2)}, \quad k=0,\ldots,(N-3)/4 \tag{3.30}$$

From the two sets of constraints (3.28) and (3.30), $G_1(z)$ is determined uniquely. To see this, note that $G_1(e^{j\omega}) = g_1(0) + 2\sum_{n=1}^{(N-1)/2} g_1(n)\cos\omega n$ is a polynomial in $x = \cos\omega$ of degree $(N-1)/2$. Since ω_k's are all distinct and $0 < \omega_k < \pi$, the constraints (3.28) and (3.30) translate into a similar set of constraints for $G_1(x)$ and $G_1'(x)$ and by simple Hermite interpolation [Dav75, page 28] $G_1(x)$ is determined uniquely. The corresponding solution $G(e^{j\omega})$ is necessarily nonnegative in the frequency region $[\pi/2,\pi]$ (Appendix). If it is nonnegative in the region $[0,\pi/2)$ as well, then it is the optimum

compaction filter with the corresponding compaction gain

$$G_{opt}(2,N) = 1 + \frac{\sum_{k=0}^{(N-3)/4} \frac{\beta_k}{2\cos(\omega_k/2)}}{r(0)} \tag{3.31}$$

If, however, $G(e^{j\omega})$ turns out to be negative at some frequencies in $[0, \pi/2)$, then it is not a valid solution and the above RHS is only an upper-bound for $G_{opt}(2,N)$. Assume that $G(e^{j\omega})$ obtained by the method is indeed nonnegative. Then it is **the unique solution!** To see this, assume there is another optimal product filter $K(z)$. Assume $K_1(z)$ is its modified polyphase component. Then, there exists a frequency ω_k among the line-spectral frequencies such that $K_1(e^{j\omega_k}) < \frac{1}{2\cos(\omega_k/2)}$. Hence the summation (3.27) for $K_1(e^{j\omega})$ is necessarily less than that for $G_1(e^{j\omega})$, resulting in contradiction. Notice finally that $H(z)$, which is an arbitrary spectral factor of the unique solution $G(z)$, is not unique.

Completion of the Optimal $G(z)$

Consider the following factorization of $G(z)$:

$$G(z) = \hat{G}_0(z)\hat{G}_1(z) \tag{3.32}$$

where $\hat{G}_0(z)$ contains the unit-circle zeros determined by the above procedure. From (3.29) we have

$$\hat{G}_0(z) = \prod_{k=0}^{\frac{N-3}{4}} (z + 2\cos(\omega_k/2) + z^{-1})^2 \tag{3.33}$$

Using the Nyquist(2) property, it is possible to determine $\hat{G}_1(z)$ and hence $G(z)$. For this, let $\hat{g}_0(n)$ and $\hat{g}_1(n)$ be the impulse responses of $\hat{G}_0(z)$ and $\hat{G}_1(z)$ respectively. The product (3.32) in $z-$domain is equivalent to the convolution in time domain. Using the convolution matrix and taking into account the symmetries, we get

$$\mathbf{g} = \mathbf{A}\hat{\mathbf{g}}_1 \tag{3.34}$$

where the vectors $\mathbf{g}, \hat{\mathbf{g}}_\mathbf{1}$ have the components

$$g_n = g(2n),\ \hat{g}_{1n} = g_1(n),\ n = 0, \ldots, (N-1)/2 \tag{3.35}$$

and $\mathbf{A}$ is obtained from the impulse response $g_0(n)$. From the Nyquist(2) property, it is clear that $\mathbf{g} = [1\ 0\ \ldots 0]^T$. Hence $\hat{\mathbf{g}}_\mathbf{1}$ is the first column of the matrix $\mathbf{A}^{-1}$. To see that $\mathbf{A}$ is invertible, it suffices to show that a unique solution to $\hat{\mathbf{g}}_\mathbf{1}$ exists for a given $g_0(n)$. For this, write the Nyquist(2) condition for $G(z)$:

$$\hat{G}_0(z)\hat{G}_1(z) + \hat{G}_0(-z)\hat{G}_1(-z) = 2 \tag{3.36}$$

The zeros of $\hat{G}_0(z)$ lie on the left half of the unit-circle. Hence the zeros of $\hat{G}_0(-z)$ lie on the right half of the unit-circle. This implies that $\hat{G}_0(z)$ and $\hat{G}_0(-z)$ are coprime. It is now easy to show that a unique solution to $\hat{G}_1(z)$ of symmetric degree less than or equal to $(N-1)/2$ exists [ADM95]. Actually, this is an efficient way of determining $\hat{G}_1(z)$ (see [ADM95] for details).

Efficient Determination of $\hat{G}_0(z)$

We will show that we can obtain $\hat{G}_0(z)$ from the singular predictor polynomial $P_1(z)$ without having to find its roots. For this, let us write $P_1(z)$ explicitly:

$$P_1(z) = dz^{-\frac{N+1}{2}} \prod_{k=0}^{\frac{N-3}{4}} (z - e^{j\omega_k})(z - e^{-j\omega_k}) = dz^{-\frac{N+1}{4}} \prod_{k=0}^{\frac{N-3}{4}} (z - 2\cos\omega_k + z^{-1}) \tag{3.37}$$

Now, consider the upsampled polynomial $P_1(z^2)$. This can be written in the form $P_1(z^2) = \pm P_0(z)P_0(-z)$, where $P_0(z)$ is a polynomial in z^{-1} of order $\frac{N+1}{2}$ with all its zeros in the left half plane. To be explicit:

$$P_0(z) = z^{-\frac{N+1}{4}} \prod_{k=0}^{\frac{N-3}{4}} (z + 2\cos(\omega_k/2) + z^{-1}) \tag{3.38}$$

Hence from (3.33) it follows that $\hat{G}_0(z) = z^{\frac{N+1}{2}} P_0^2(z)$. Therefore, given the singular predictor polynomial $P_1(z)$, one can apply a continuous-time spectral factorization algorithm [Bau55] to $P_1(z^2)$ to obtain $P_0(z)$ and therefore $\hat{G}_0(z)$. Since $G(z)$ can be determined from $\hat{G}_0(z)$, we observe that there is no need to find the roots of $P_1(z)$!

Spectral Factorization

To find the compaction filter $H(z)$, we need to spectrally factorize $G(z)$. It is clear that we can write $H(z)$ as

$$H(z) = H_0(z)H_1(z)$$

where $H_0(z)$ and $H_1(z)$ are the spectral factors of $\hat{G}_0(z)$ and $\hat{G}_1(z)$ respectively. We can deduce $H_0(z)$ immediately: $H_0(z) = P_0(z)$. Hence all we need to do is to determine $H_1(z)$ which is of order $\frac{N-1}{2}$. This can be done by a discrete-time spectral factorization of $\hat{G}_1(z)$ [Ngu92]. Although the phase of the compaction filter is immaterial for the compaction gain, it is important in the design of an optimal orthonormal filter bank for subband coding as we saw in Sec. 2.2.4. For some applications like image coding, linear-phase property might be important. Although it is not possible to have linear-phase compaction filter in the two-channel case [Vai93], one can achieve close-to-linear-phase response by a careful grouping of the roots of $\hat{G}_1(z)$.

The case where $\frac{N-1}{2}$ is even can be treated in a very similar manner. In this case, we use the singular polynomial $P_{-1}(z)$ corresponding to $c = -1$ and one of the line-spectral frequencies is 0, that is, $z = 1$ is a root of $P_{-1}(z)$. The resulting product filter $G(e^{j\omega})$ continues to be nonnegative in $[\pi/2, \pi]$. We skip the details and give the summary of the algorithm for both cases.

Summary of the Analytical Method

Given the autocorrelation sequence $r(n)$, $n = 0, \ldots, N$, where N is odd, let $m = (N-1)/2$. First obtain the sequence $\psi_x(n)$, $n = 0, \ldots, m$ using the relations (3.22). **If this sequence is positive definite**, then

Step 1. Calculate $A_m(z)$, the optimum predictor polynomial of order m, correspond-

ing to the sequence $\psi_x(n)$ and obtain $P_c(z)$ from $P_c(z) = A_m(z)+cz^{-(m+1)}A_m(z^{-1})$, where $c = 1$ if m is odd, and $c = -1$ otherwise.

Step 2. Obtain the spectral factor, $P_0(z)$ of $P_c(z^2)$ using a continuous time spectral factorization algorithm and determine $\hat{G}_0(z) = z^{(m+1)}P_0^2(z)$.

Step 3. Calculate $\hat{G}_1(z)$ using (3.34) or (3.36) and find its spectral factor $H_1(z)$. The optimum compaction filter is $H(z) = P_0(z)H_1(z)$.

See our webpage [htt] for a MATLAB program that implements the algorithm.

Decorrelation in Optimal Subband Coding

Let us form a two-channel orthonormal filter bank by letting the first filter be the optimal FIR compaction filter $H(z)$ designed above and by having the second filter as [Vai93]

$$F(z) = z^{-N}\tilde{H}(-z) \tag{3.39}$$

Let $S_{x_0x_1}(z)$ be the cross spectral density of the subband signals after decimation. Then we have

$$\begin{aligned} S_{x_0x_1}(z) &= \left[S_{xx}(z)H(z)\tilde{F}(z)\right]_{\downarrow 2} \\ &= \left[z^N S_{xx}(z)H(z)H(-z)\right]_{\downarrow 2} \\ &= \left[z^N S_{xx}(z)P_c(z^2)H_1(z)H_1(-z)\right]_{\downarrow 2} \\ &= \left[z^N S_{xx}(z)\hat{G}_1(z)\right]_{\downarrow 2} P_c(z) \end{aligned} \tag{3.40}$$

Hence we have

$$S_{x_0x_1}(e^{j\omega_k}) = 0, \quad k = 0, \ldots, m, \tag{3.41}$$

where ω_k's are the line-spectral frequencies (see Remarks after Theorem 1). This is the form of decorrelation that takes place in optimal subband coding with FIR filters.

Case Where $\psi_x(n)$ is Negative Definite

From our developments for the positive definite case, and using the sequence $-\psi_x(n)$, it can be proven that the optimum compaction filter is $H(z) = \hat{H}(-z)$ where $\hat{H}(z)$ is the optimum compaction filter for the positive definite sequence $\hat{\psi}_x(n) = -\psi_x(n)$. However, it is easier to see this directly by looking at the objective in time domain: $\sigma_y^2 = r(0) + 2\sum_{n=1}^{N} g(n)r(n)$. First note that $\hat{\psi}_x(n)$ corresponds to the autocorrelation sequence $\hat{r}(n) = -r(n),\ n \neq 0$. Let $g(n)$ and $\hat{g}(n)$ be the product filter coefficients for $H(z)$ and $\hat{H}(z)$ respectively. The objective is then to maximize $\sum_{n=1}^{N} -g(n)\hat{r}(n)$. This has the solution $-g(n) = \hat{g}(n),\ n \neq 0$. Hence we have $G(z) = \hat{G}(-z)$ and therefore $H(z) = \hat{H}(-z)$.

Example 1: AR(1) process. Let the input process be $AR(1)$ with the autocorrelation sequence $r(n) = \rho^n,\ 0 < \rho < 1$. This is also called Markov-1 process and is a good model for many of the practical signals including images and speech signals [Jai89]. Let the compaction filter order be $N = 3$. Then, $m = 1$ which is odd. We have $\psi_x(0) = 2\rho$ and $\psi_x(1) = \rho(1+\rho^2)$. The Hermitian Toeplitz matrix corresponding to $\{\psi_x(n),\ n = 0, 1\}$ is

$$\mathbf{P} = \rho \begin{bmatrix} 2 & 1+\rho^2 \\ 1+\rho^2 & 2 \end{bmatrix} \tag{3.42}$$

which is positive definite. Hence we can apply the analytical method:

Step 1. Running the Levinson recursion, we have: $A_1(z) = 1 - \frac{1+\rho^2}{2}z^{-1}$ and using $c = 1$ (m is odd) we have: $P_1(z) = 1 - (1+\rho^2)z^{-1} + z^{-2}$.

Step 2. By straightforward calculation $P_0(z) = 1 + \sqrt{3+\rho^2}z^{-1} + z^{-2}$ and $\hat{G}_0(z) = (z + \sqrt{3+\rho^2} + z^{-1})^2$.

Step 3. Using the Nyquist(2) constraint we find

$$\hat{G}_1(z) = -\frac{1}{2(3+\rho^2)^{3/2}}(z - 2\sqrt{3+\rho^2} + z^{-1}) \tag{3.43}$$

It is readily verified that $\hat{G}_1(e^{j\omega}) \geq 0,\ \forall\omega$ for all values of ρ. The spectral factor

of $\hat{G}_1(z)$ turns out to be

$$\frac{1}{\sqrt{2}(3+\rho^2)^{3/4}}(a+bz^{-1}) \tag{3.44}$$

where

$$a=\sqrt{\sqrt{3+\rho^2}+\sqrt{2+\rho^2}} \quad \text{and} \quad b=-\sqrt{\sqrt{3+\rho^2}-\sqrt{2+\rho^2}} \tag{3.45}$$

Step 4. The optimum compaction filter is

$$\begin{aligned} H(z) &= P_0(z)H_1(z) \\ &= \frac{1}{\sqrt{2}(3+\rho^2)^{3/4}}\left(a+(b+a\sqrt{3+\rho^2})z^{-1}+(a+b\sqrt{3+\rho^2})z^{-2}+bz^{-3}\right) \end{aligned} \tag{3.46}$$

The product filter is

$$G(z)=\frac{1}{2(3+\rho^2)^{3/2}}(-z^3+3(2+\rho^2)z+2(3+\rho^2)^{3/2}+3(2+\rho^2)z^{-1}-z^{-3}) \tag{3.47}$$

If $-1<\rho<0$, then the optimum compaction filter is $H(-z)$. The optimum compaction gain for both cases is

$$G_{opt}(2,3)=1+\frac{2|\rho|}{\sqrt{3+\rho^2}} \tag{3.48}$$

See Table 3.1 for the numerical values of the filter coefficients and the compaction gains for various values of ρ. We have found that the analytical method is successful for any filter order N for AR(1) processes.

Example 2: MA(1) process. Let $N=3$, $r(0)=1$, $r(1)=\rho>0$, and $r(n)=0$, $n>1$. The sequence $\psi_x(n)$ is therefore $\psi_x(0)=2\rho$, $\psi_x(1)=\rho$.

$$\mathbf{P}=\rho\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix} \tag{3.49}$$

$\rho = 0.1$

n	analytical method	window method	linear programming
0	0.5494144350	0.6940928372	0.5839818982
1	0.7789293967	0.7136056607	0.7658293099
2	0.2470689810	0.0680132766	0.2140666953
3	-0.1742690225	-0.0661535225	-0.1632362113
compaction gain	1.1153	1.1078	1.1151

$\rho = 0.5$

n	analytical method	window method	linear programming
0	0.5308991349	0.6817974052	0.5693221037
1	0.7963487023	0.7258587819	0.7821354608
2	0.2411149862	0.0663296736	0.2047520302
3	-0.1607433241	-0.0623033026	-0.1490404954
compaction gain	1.5547	1.5283	1.5537

$\rho = 0.9$

n	analytical method	window method	linear programming
0	0.4938994371	0.6550553981	0.5605331011
1	0.8279263239	0.7510864372	0.8017336546
2	0.2281902949	0.0620169861	0.1699982390
3	-0.1361269173	-0.0540877314	-0.1188544843
compaction gain	1.9222	1.9118	1.9207

Table 3.1: The optimum compaction filter coefficients $h(n)$ and the corresponding compaction gains for AR(1) process with $\rho = 0.1, 0.5$, and 0.9. Here filter order is $N = 3$ and the number of channels is $M = 2$.

which is positive definite. Hence applying the algorithm, we find:

$$\begin{aligned} A_1(z) &= 1 - \frac{1}{2}z^{-1}, \quad c = 1, \quad P_1(z) = 1 - z^{-1} + z^{-2} \\ P_0(z) &= 1 + 2\cos(\pi/6)z^{-1} + z^{-2} = 1 + \sqrt{3}z^{-1} + z^{-2} \\ \hat{G}_0(z) &= (z + \sqrt{3} + z^{-1})^2, \quad \hat{G}_1(z) = -\frac{\sqrt{3}}{18}(z - 2\sqrt{3} + z^{-1}) \\ H_1(z) &= 3^{-3/4}2^{-1/2}(\sqrt{\sqrt{3} + \sqrt{2}} - \sqrt{\sqrt{3} - \sqrt{2}}z^{-1}) \end{aligned} \tag{3.50}$$

and the compaction filter is

$$\begin{aligned} H(z) &= 3^{-3/4}2^{-1/2}(\sqrt{\sqrt{3}+\sqrt{2}}+(\sqrt{3+\sqrt{6}}-\sqrt{\sqrt{3}-\sqrt{2}})z^{-1} \\ &\quad +(\sqrt{\sqrt{3}+\sqrt{2}}-\sqrt{3-\sqrt{6}})z^{-2}-\sqrt{\sqrt{3}-\sqrt{2}}z^{-3}) \end{aligned} \tag{3.51}$$

The product filter is

$$G(z) = -\frac{\sqrt{3}}{18}z^3 + \frac{\sqrt{3}}{3}z + 1 + \frac{\sqrt{3}}{3}z^{-1} - \frac{\sqrt{3}}{18}z^{-3} \tag{3.52}$$

If $\rho < 0$, then the optimal filter is $H(-z)$. The optimum compaction gain for both cases is

$$G_{opt}(2,3) = 1 + \frac{2}{\sqrt{3}}|\rho| \tag{3.53}$$

Example 3: MA(1) process, arbitrary order N. Following the steps of the algorithm, we have

$$P_c(z) = 1 - z^{-1} + z^{-2} - \ldots + (-1)^{\frac{N+1}{2}}z^{-\frac{N+1}{2}} \tag{3.54}$$

If $\frac{N-1}{2}$ is odd, then the zeros of $P_1(z)$ are

$$e^{\pm j\omega_k}, \quad \omega_k = (2k-1)\frac{2\pi}{N+3}, \; k = 1, \ldots, (N+1)/4 \tag{3.55}$$

Therefore, the roots of $P_0(z)$, hence the unit-circle zeros of the optimum compaction filter $H(z)$, are

$$e^{\pm j\Omega_k}, \quad \Omega_k = \pi - (2k-1)\frac{\pi}{N+3}, \quad k = 1, \ldots, \frac{N+1}{4} \tag{3.56}$$

Similarly, if $\frac{N-1}{2}$ is even, the unit-circle zeros of the optimum compaction filter $H(z)$ are

$$\pi, \quad e^{\pm j\Omega_k}, \quad \Omega_k = \pi - 2k\frac{\pi}{N+3}, \; k = 1, \ldots, \frac{N-1}{4} \tag{3.57}$$

The rest of the procedure involves spectral factorization and it is not easy to see what

$H_1(z)$ will be in closed form. However, we note that the algorithm successfully finds the optimum compaction filter for any order N. Table 3.2 shows the compaction filters and the corresponding compaction gains for various filter orders. The optimum compaction filter for $\rho < 0$ is $H(-z)$. Note that the filters do not depend on the value of ρ but only on the sign. The optimum compaction gains, on the other hand, depend on ρ.

n	N = 3	N = 9	N = 15	N = 21
0	0.5502267080	0.3472380509	0.2619442448	0.2135948251
1	0.7781380728	0.7212669193	0.6444985282	0.5837095513
2	0.2473212614	0.5313628729	0.6197371546	0.6522949264
3	-0.1748825411	-0.0301418144	0.1178983343	0.2237822545
4		-0.2357012104	-0.2498547909	-0.2185003959
5		0.0008621669	-0.1127984531	-0.1849242462
6		0.1250275869	0.1367316336	0.1009864921
7		-0.0141611881	0.0800586128	0.1357184335
8		-0.0608205190	-0.0879123348	-0.0574793351
9		0.0292806975	-0.0512638394	-0.0996883819
10			0.0616834351	0.0386787054
11			0.0272577935	0.0732876341
12			-0.0441106561	-0.0298852666
13			-0.0065141401	-0.0529392251
14			0.0275486991	0.0255464898
15			-0.0111966480	0.0362662187
16				-0.0230508869
17				-0.0216340483
18				0.0206081274
19				0.0077883397
20				-0.0156868986
21				0.0057402527
$\rho = 0.1$	1.1155	1.1244	1.1260	1.1266
comp. $\rho = 0.3$	1.3464	1.3732	1.3781	1.3798
gains $\rho = 0.5$	1.5774	1.6220	1.6301	1.6330

Table 3.2: Compaction filter coefficients and corresponding gains for MA(1) processes, for $M = 2$.

Example 4: KLT. If $N = 1$, then the algorithm yields $H(z) = \frac{1}{\sqrt{2}}(1+z^{-1})$ if $r(1) > 0$ and $H(-z) = \frac{1}{\sqrt{2}}(1-z^{-1})$ if $r(1) < 0$. Notice that these correspond to the two-channel transform coder which is known to be fixed. The corresponding compaction gain is

$G_{opt}(2,1) = 1 + \frac{|r(1)|}{r(0)}$. It is also true that the above filters and the corresponding compaction gains are optimal for any psd and for any number of channels M. Hence

$$G_{opt}(M,1) = 1 + \frac{|r(1)|}{r(0)}, \quad M \geq 2 \tag{3.58}$$

If $r(m)$ is maximum of all $r(n)$ where n is not a multiple of M, then one can achieve the compaction gain of $1 + \frac{|r(m)|}{r(0)}$ by using the filter $\frac{1}{\sqrt{2}}(1 + z^{-m})$ if $r(m) > 0$ and the filter $\frac{1}{\sqrt{2}}(1 - z^{-m})$ if $r(m) < 0$.

Case Where $\psi_x(n)$ is Semidefinite.

Assume that $\psi_x(n)$ is positive semidefinite. Then there exists an integer $P < (N-1)/2$ such that $\{\psi_x(n),\ n = 0, 1, ..., P\}$ is positive definite and $\{\psi_x(n),\ n = 0, 1, ..., P+1\}$ is only positive semidefinite. Then we can replace $(N-1)/2$ in the above arguments with P and write the objective (3.27) in terms of $P+1$ corresponding line-spectral frequencies. This enables us to determine a product filter of symmetric order $2P+1 < N$. If this resulting filter is nonnegative, then we have found the unique minimum symmetric order product filter that is optimum among the filters of symmetric order less than or equal to N! The case where $\psi_x(n)$ is negative semidefinite is similar; the details are omitted.

Example 5: Case where $\psi_x(n)$ is positive semidefinite. Let $N = 3$, $r(0) = 1$ and $r(1) = r(3) = \rho > 0$. Then, $\psi_x(0) = \psi_x(1) = 2\rho$. The associated Toeplitz matrix is

$$\mathbf{P} = 2\rho \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{3.59}$$

which is positive semidefinite and singular. The number P is 0 in this case and the objective (3.27) is $1 + 2\rho G_1(e^{j0})$. By letting $G_1(e^{j0}) = \frac{1}{2}$, the product filter $G(z)$ of symmetric order 1 can easily be seen to be $\frac{1}{2}z + 1 + \frac{1}{2}z^{-1}$, and it is readily verified that $G(e^{j\omega}) \geq 0$. In fact, this is the KLT solution with the compaction filter $H(z) = \frac{1}{\sqrt{2}}(1 + z^{-1})$. The corresponding optimum compaction gain is $1 + \rho$. No 3rd order solution can achieve better gain than this.

3.3.3 Characterization of Processes for Which the Analytical Method is Applicable for all N.

For the analytical method to be applicable for all N, the sequence $\psi_x(n)$ has to be positive or negative definite for all N. The sequence $\psi_x(n)$ is positive definite for all N, if and only if $\Psi_x(e^{j\omega})$ is not a line-spectrum and $\Psi_x(e^{j\omega}) \geq 0$. Using (3.20), this is true if and only if $S_{xx}(e^{j\omega})$ is not a line-spectrum and

$$S_{xx}(e^{j\omega}) \geq S_{xx}(e^{j(\pi-\omega)}),\ \omega \in [0, \pi/2] \tag{3.60}$$

We will say that the process is 'low-pass' if its psd satisfies the above condition. A nonincreasing psd is an example of this. However, a psd may not be nonincreasing but may still be low-pass. In the ideal case, the optimum compaction filter for that type of process is the ideal half-band low-pass filter [DMS92, TG95, Uns93b]. For the case where $\psi_x(n)$ is negative definite for all N, the preceding is replaced with $S_{xx}(e^{j\omega}) \leq S_{xx}(e^{j(\pi-\omega)}),\ \omega \in [0, \pi/2]$. This type of process will be called 'high-pass' since the ideal half-band high-pass filter is optimum for such a process. Notice that for the algorithm to be applicable for a particular N, it is only necessary that $\psi_x(n),\ n = 0, \ldots, (N-1)/2$ is positive or negative definite. For a small order N, this corresponds to a much broader class than that of low-pass and high-pass processes.

Cases Where the Algorithm Fails

Assume that the process is such that the sequence $\{\psi_x(n),\ n = 0, \ldots, (N-1)/2\}$ is positive definite and therefore the algorithm is applicable for the filter order N. Assume, however, that one of the line-spectral frequencies ω_k is close to π. The algorithm will require $e^{j(\pi-\omega_k/2)}$ to be a zero of $G(z)$. Hence $G(e^{j\omega})$ will have a zero close to $\pi/2$. But this may be impossible if the order N is low. To see this, note that $G(e^{j\pi/2}) = 1$ from the Nyquist(2) property and therefore requiring $G(e^{j\omega})$ to have a zero close to the frequency $\pi/2$ is the same as requiring a narrow transition band for $G(e^{j\omega})$ which is impossible if the order is not sufficiently high. One can, however, increase the filter

order to overcome the problem.

Example 6. Let $N = 3$, and $r(n) = \cos\omega_1 n$, $\omega_1 \in [0, \pi/2)$. Hence $\psi_x(0) = 2\cos\omega_1$, $\psi_x(1) = \cos 3\omega_1 + \cos\omega_1$, and $\psi_x(n)$ is positive definite. Using the algorithm, we find $\hat{G}_0(z) = (z + 2\cos\omega_1 + z^{-1})^2$ from which it follows that

$$\hat{G}_1(z) = -\frac{1}{16\cos^3\omega_1}(z - 4\cos\omega_1 + z^{-1}) \tag{3.61}$$

This has single unit-circle zeros if $\omega_1 \in (\pi/3, \pi/2)$ and therefore $\hat{G}_1(e^{j\omega})$ is not non-negative. Hence the algorithm fails if the impulse is within $\pi/6$ neighborhood of $\pi/2$. We have designed optimum compaction filters for the above autocorrelation sequence using LP for various values of ω_1. We have observed that the optimum compaction filters agree with the above analytical solution if $\omega_1 \in (0, \pi/3]$. For the complementary case of $\omega_1 \in (\pi/3, \pi/2)$ where the analytical method fails for $N = 3$, LP yields the solution $G(z) = -\frac{1}{2}z^3 + 1 - \frac{1}{2}z^{-3}$, *regardless of the exact value of* ω_1. The factors $\hat{G}_0(z)$ and $\hat{G}_1(z)$ of $G(z)$ are

$$\hat{G}_0(z) = (z + 2\cos(\pi/3) + z^{-1})^2 \quad \text{and} \quad \hat{G}_1(z) = -\frac{1}{16\cos^3(\pi/3)}(z - 4\cos(\pi/3) + z^{-1}) \tag{3.62}$$

This is the same as the previous solution except that ω_1 in the previous solution is replaced with a constant value equal to $\pi/3$.

As another example, let us fix $\omega_1 = 2\pi/5 > \pi/3$, and find the optimal FIR compaction filter of order 5. The corresponding product filter is $G(z) = \frac{1}{2}z^5 + 1 + \frac{1}{2}z^{-5}$ and the compaction gain is $G_{opt}(2, 5) = 2$ which is the largest possible gain for $M = 2$! Since the process is line-spectral, this is not surprising. *The important point here is that while the algorithm is not successful for the filter order* 3, *it is successful for a higher order* 5.

Example 7: Case where the process is multiband. Finally we will consider an example in which the input is not low-pass or high-pass but rather it is of multiband nature. Let $r(0) = 1, r(1) = \frac{1}{10}, r(2) = 0$, and $r(3) = -\frac{1}{4}$. The sequence $\psi_x(n)$ is positive definite for $N = 3$ so that the algorithm is applicable. There is more than

one way to extrapolate this sequence and find the corresponding psd. For example, one can consider MA(3), AR(3), or line-spectra(4). In all three cases, we have verified that the psd is neither low-pass nor high-pass. Rather it is of multi-band nature. Applying the algorithm steps we have $\hat{G}_0(z) = (z + \frac{1}{\sqrt{2}} + z^{-1})^2$ from which it follows that $\hat{G}_1(z) = -\sqrt{2}(z - \sqrt{2} + z^{-1})$. This has single unit-circle zeros! Hence $\hat{G}_1(e^{j\omega})$ is not nonnegative and therefore $G(e^{j\omega}) = \hat{G}_0(e^{j\omega})\hat{G}_1(e^{j\omega})$ is not nonnegative either. The algorithm halts because $\hat{G}_1(e^{j\omega})$ cannot be spectrally factorized.

3.4 Window Method

In this section we will describe a new method to design FIR compaction filters. The method is applicable for arbitrary filter order N, arbitrary number of channels M, and for any given psd (including complex and multiband spectra). The technique is quite simple while the resulting compaction gains are very close to the optimum ones especially for high filter orders.

A common practice in filter design is to approximate ideal filter responses by windowing their impulse responses. Consider the ideal compaction filter design: for each $\omega \in [0, 2\pi/M)$, let $S_{xx}(e^{j(\omega + i_0 \frac{2\pi}{M})})$ be the maximum of the set

$$\{S_{xx}(e^{j(\omega + i\frac{2\pi}{M})}),\ i = 0, \ldots, M-1\} \tag{3.63}$$

Then

$$H_i(e^{j(\omega + i_0 \frac{2\pi}{M})}) = \sqrt{M} \quad \text{and} \quad H_i(e^{j(\omega + i\frac{2\pi}{M})}) = 0 \quad \text{for} \quad i \neq i_0 \tag{3.64}$$

Let $h_i(n)$ be the impulse response of $H_i(e^{j\omega})$, and consider

$$h(n) = w(n)h_i(n) \tag{3.65}$$

for a given finite length window $w(n)$. Let $H(e^{j\omega})$ be the FT of $h(n)$. Then $G(e^{j\omega}) = |H(e^{j\omega})|^2$ is no longer Nyquist(M). Instead of windowing $h_i(n)$, let us try to window the coefficients of the product filter: $g(n) = w(n)g_i(n)$. Here $g_i(n)$ is the impulse

response of $G_i(e^{j\omega}) = |H_i(e^{j\omega})|^2$. Then $G(e^{j\omega})$ is Nyquist(M) but it may no longer be nonnegative. The nonnegativity can also be assured by constraining the FT of $w(n)$ to be nonnegative. A compaction filter can then be successfully obtained by spectrally factorizing $G(e^{j\omega})$. This can be considered as the approximation of the ideal compaction filter response.

In this section we extend this idea to design compaction filters that perform better than the above ad hoc windowing of ideal compaction filters. We will replace $g_i(n)$ with a periodic sequence $f_L(n)$ which will be determined by applying the ideal design algorithm at L uniform DFT frequencies. If $L = \infty$, then we have $f_L(n) = g_i(n)$, and the above ad hoc method results as a special case. It turns out that the experimentally optimum value of L for the best compaction gain is $L = M\lceil 2N/M \rceil$ (see Sec. 3.4.2).

3.4.1 Derivation of the Window Method

To formalize the above ideas, let us write the product filter coefficients $g(n)$ in the form

$$g(n) = w(n) f_L(n), \tag{3.66}$$

where $w(n)$ has the same length as $g(n)$, namely $2N + 1$ and $f_L(n)$ is a periodic sequence with period $L = KM \geq 2N$ for some K (see Fig. 3.4). Let $W(e^{j\omega})$ be

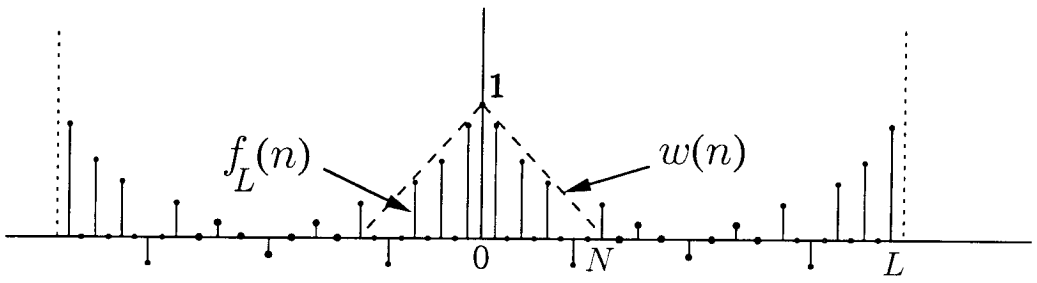


Fig. 3.4: Decomposition of $g(n)$ as $w(n)f_L(n)$ where $W(e^{j\omega}) \geq 0$ and $F_L(k) \geq 0$.

the FT of $w(n)$ and $F_L(k)$ be the Fourier series coefficients (FSC) of $f_L(n)$, that is $F_L(k) = \sum_{n=0}^{L-1} f_L(n) W_L^{kn}$, $W_L = e^{-j2\pi/L}$. The first period of $F_L(k)$ is just the DFT of the first period of $f_L(n)$. We have the following observation:

Lemma 1. Consider (3.66). If (i) $w(0) = 1$, (ii) $W(e^{j\omega}) \geq 0$, (iii) $F_L(k) \geq 0$, (iv) $f_L(n)$ is Nyquist(M), then $G(z)$ is the product filter of a valid compaction filter. That is, $g(Mn) = \delta(n)$ and $G(e^{j\omega}) \geq 0$.

Proof. It is readily verified that $G(e^{j\omega}) = \frac{1}{L}\sum_{k=0}^{L-1} F_L(k)W(e^{j(\omega-\frac{2\pi}{L}k)})$. Since $F_L(k) \geq 0$ and $W(e^{j\omega}) \geq 0$, it follows that $G(e^{j\omega}) \geq 0$. If $f_L(n)$ is Nyquist(M) then so is $g(n)$ because $g(Mn) = w(Mn)f_L(Mn) = \delta(n)$. ∎

Assume the conditions of the lemma hold so that $G(z)$ is the product filter of a valid compaction filter. If $w(n)$ and L is fixed, what is the best $f_L(n)$ that maximizes the compaction gain? To answer the question first note:

Lemma 2. A periodic sequence $f_L(n)$ with period $L = KM$ is Nyquist(M), that is, $f_L(Mn) = \delta_K(n)$, if and only if its FSC $F_L(k)$ satisfy the following:

$$\sum_{i=0}^{M-1} F_L(k + iK) = M, \quad k = 0, \ldots, K-1 \tag{3.67}$$

Proof. Let us find the FSC $Y_K(k)$, of the decimated sequence $y_K(n) = f_L(Mn)$: $Y_K(k) = \sum_{n=0}^{K-1} f_L(Mn)W_K^{kn}$. This can be written as

$$Y_K(k) = \frac{1}{M}\sum_{i=0}^{M-1}\sum_{j=0}^{K-1} F_L(j + iK)\frac{1}{K}\sum_{n=0}^{K-1} W_K^{(k-j)n} \tag{3.68}$$

Using $\frac{1}{K}\sum_{n=0}^{K-1} W_K^{mn} = \delta_K(m)$ we have $Y_K(k) = \frac{1}{M}\sum_{i=0}^{M-1} F_L(k + iK)$. The FSC of $\delta_K(n)$ are all 1. Hence $f_L(Mn) = \delta_K(n)$ if and only if $\frac{1}{M}\sum_{i=0}^{M-1} F_L(k + iK) = 1$, $\forall k = 0, \ldots, K-1$. ∎

To obtain the best $f_L(n)$, let $\hat{r}(n) = w^*(n)r(n)$ and let $\hat{S}_L(k)$ be the FSC of its periodic expansion $\hat{r}_L(n)$. For simplicity assume that $L > 2N$. The objective (3.10) becomes

$$\sigma_y^2 = \sum_{n=0}^{L-1} f_L(n)\hat{r}_L^*(n) = \frac{1}{L}\sum_{k=0}^{L-1} F_L(k)\hat{S}_L(k) \tag{3.69}$$

Both $F_L(k)$ and $\hat{S}_L(k)$ are real. Now to incorporate the Nyquist(M) constraint, we

write the preceding as

$$\sigma_y^2 = \frac{1}{L} \sum_{k=0}^{K-1} \sum_{i=0}^{M-1} F_L(k+iK) \hat{S}_L(k+iK) \tag{3.70}$$

For a fixed k, let $\hat{S}_L(k+i_0K)$ be the maximum of the set

$$\{\hat{S}_L(k+iK),\ i=0,\ldots,M-1\} \tag{3.71}$$

Then by (3.67), and noting that $F_L(k) \geq 0$, the objective (3.70) is maximized if we assign

$$F_L(k+i_0K) = M, \quad \text{and} \quad F_L(k+i_lK) = 0,\ l = 1,\ldots,M-1. \tag{3.72}$$

Repeating the process for each $k = 0,\ldots,K-1$, the FSC of the best $f_L(n)$ is determined. The procedure is illustrated in Fig. 3.5. The sequence $f_L(n)$ is just the inverse

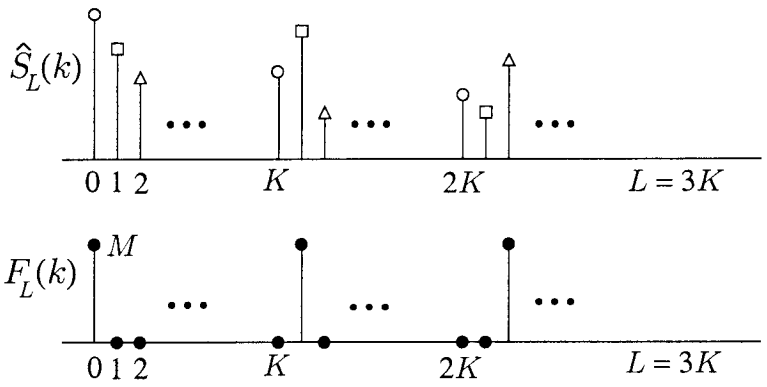


Fig. 3.5: The procedure to find $F_L(k)$: $\hat{S}_L(0)$ is maximum among $\{\hat{S}_L(iK)\}$, hence $F_L(0) = M$, $F_L(lK) = 0, l \neq 0$. $\hat{S}_L(1+K)$ is maximum among $\{\hat{S}_L(1+iK)\}$, hence $F_L(1+K) = M$, $F_L(1+lK) = 0, l \neq 1$, and so on.

DFT of $F_L(k)$:

$$f_L(n) = \frac{1}{L} \sum_{k=0}^{L-1} F_L(k) W_L^{-nk} \tag{3.73}$$

Summary of the Window Algorithm

Assume a window $w(n)$ of the same length as $g(n)$ with nonnegative FT has been chosen. Let $L = KM > 2N$. Then the algorithm steps are

Step 1. Calculate $\hat{S}_L(k)$, the L-point DFT of the conjugate-symmetric sequence $\hat{r}(n) = w^*(n)r(n)$ (same as the FSC of the periodical expansion $\hat{r}_L(n)$ of $\hat{r}(n)$).

Step 2. For each $k = 0, \ldots, K-1$, determine the index i_0 for which $\hat{S}_L(k + i_0 K)$ is maximum, and assign $F_L(k + i_0 K) = M$ and $F_L(k + i_l K) = 0$, $l = 1, \ldots, M-1$.

Step 3. Calculate $f_L(n)$ by the inverse DFT. We need only to determine $f_L(n)$ for $n = 1, \ldots, N$.

Step 4. Form the product filter $g(n) = w(n) f_L(n)$ and spectrally factorize it to find $H(z)$.

Real Case. If the input is real, the above algorithm can be modified to produce real-coefficient compaction filters. Consider the set $\{\hat{S}_L(k + iK),\ i = 0, \ldots, M-1\}$ for each $k = 0, \ldots, K-1$. Since $\hat{S}_L(k) = \hat{S}_L(L-k)$ if the process is real, this set is equivalent to $\{\hat{S}_L(L - k - iK) = \hat{S}_L(K - k + K(M-1-i)),\ i = 0, \ldots, M-1\}$. Hence in the comparison, we need to consider only $k = 0, \ldots, P$ where $P = \frac{K}{2}$ if K is even, and $P = \frac{K-1}{2}$ if it is odd. Let $\hat{S}_L(k + i_0 K)$ be the maximum of this set for each $k = 0, \ldots, P$. We need to be careful in the assignments. The symmetric frequencies may end up in the same set and we cannot assign different values to them. There are two cases to consider:

i) the index $L - k - i_0 K$ is among the set $\{k + iK,\ i = 0, \ldots, M-1\}$,

ii) it is not.

The first case happens if and only if $2k \ mod \ K = 0$. This happens if $k = 0$ or $k = \frac{K}{2}$. We assign $F_L(k + i_0 K) = F_L(L - k - i_0 K) = \frac{M}{2}$ if $k + i_0 K \neq \frac{L}{2}$, and $F_L(k + i_0 K) = M$ if $k + i_0 K = \frac{L}{2}$. In the second case, we assign $F_L(k + i_0 K) = F_L(L - k - i_0 K) = M$ if $k + i_0 K \neq 0$ and $F_L(k + i_0 K) = M$ if $k + i_0 K = 0$. In either case, we set the remaining values in the set $\{F_L(k + iM)\}$ to zeros. This will maximize the objective (3.70), and $f_L(n)$ calculated by the inverse DFT is the best sequence and it is real.

Summary of the Window Algorithm for the Real Case

Assume a real symmetric window $w(n)$ of order N, with nonnegative FT is given. Let $L = KM > 2N$ as before. Let P be as explained above. Then Step 2 of the previous algorithm should be replaced by the following two steps:

Step 2.1. For each $k = 0, \ldots, P$, determine the index i_0 for which $\hat{S}_L(k + i_0K)$ is maximum.

Step 2.2. If $k + i_0K = 0$ or $k + i_0K = \frac{L}{2}$, then set $F_L(k + i_0K) = M$, else if $k = 0$ or $k = \frac{K}{2}$, then set $F(k + i_0K) = F(L - k - i_0K) = \frac{M}{2}$, else, set $F(k + i_0K) = F(L - k - i_0K) = M$. Set the remaining to zeros.

Optimization of the Window

The algorithm produces very good compaction gains especially when the filter order is high as we shall demonstrate shortly. However, one can get better compaction gains by optimizing the window $w(n)$. Consider the representation (3.66) again and let $w(n)$ and $f_L(n)$ satisfy the conditions of Lemma 1. If we fix $f_L(n)$, what is the best window $w(n)$? The objective (3.10) can be written as

$$\sigma_y^2 = \int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega}) W(e^{j\omega}) \frac{d\omega}{2\pi} \tag{3.74}$$

where $W(e^{j\omega})$ is the FT of $w(n)$ and $\hat{S}_{xx}(e^{j\omega})$ is the FT of $f^*(n)r(n)$ where $f(n)$ is one period of $f_L(n)$ centered at $n = 0$. Let $W(e^{j\omega}) = |A(e^{j\omega})|^2$, where $A(z) = \sum_{n=0}^{N} a(n)z^{-n}$ is the spectral factor of $W(z)$. The only constraint on $A(e^{j\omega})$ is that it has to have unit energy in view of

$$w(0) = \int_{-\pi}^{\pi} |A(e^{j\omega})|^2 \frac{d\omega}{2\pi} = 1 \tag{3.75}$$

Let $\mathbf{P}$ be the $(N+1)\times(N+1)$ Hermitian Toeplitz matrix corresponding to the sequence $\left\{f_L^*(n)r(n)\right\}_0^N$. Then, by Rayleigh's principle [HJ85], (3.74) is maximized if $A(z)$ is the maximal eigenfilter of $\mathbf{P}$. The corresponding compaction gain is

$$\lambda_{max}\left\{f_L^*(n)r(n)\right\}_0^N / \sigma_x^2 \tag{3.76}$$

Corollary. A lower bound on the compaction gain. Let $f_L(n)$ be any Nyquist(M) sequence with nonnegative FSC. Assume $L > N$. Then,

$$G_{opt}(M,N) \geq \lambda_{max}\left\{f_L^*(n)r(n)\right\}_0^N \tag{3.77}$$

To see this note that $g(n) = w(n)f_L(n)$ achieves that bound by choosing $w(n)$ as the optimum window for the sequence $f_L(n)$. If we replace $f_L(n)$ by a positive definite Nyquist(M) sequence $f(n)$ of order N, the inequality continues to be valid because $w(n)f(n)$ is still a product filter of a valid compaction filter. To see this, note that the sequence $f(n)$ can be extended to an infinite sequence (e.g., using autoregressive extrapolation) such that its FT is nonnegative. Hence the product $w(n)f(n)$ has nonnegative FT. The Nyquist(M) property of the product follows from that of $f(n)$.

We have described how to optimize $w(n)$ given $f_L(n)$, and vice versa. It is reasonable to expect that one can iterate and obtain better compaction gains at each stage. We have observed that this is not the case. We started with a triangular window and found that $f_L(n)$ did not change after the re-optimization of the window. Notice that the use of an initial window is not necessary if one is willing to optimize the window after finding $f_L(n)$. However, in most of the design examples we considered, using an initial window with nonnegative FT (in particular, the triangular window) and then re-optimizing the window resulted in better compaction gains. A MATLAB program that implements the window method can be found at our webpage [htt]. Here is a simple example to illustrate how the window method works:

Example 8: MA(1) process. Let $N = 5$, $M = 4$, $r(0) = 1$, $r(1) = \rho$, and $r(n) = 0$, $n > 1$. Assume the process is real so that $r(-n) = r(n)$. Let the window be triangular, i.e.,

$$w(n) = \begin{cases} 1 - \frac{|n|}{6}, & n = 0, \pm 1, \ldots, \pm 5 \\ 0, & \text{elsewhere.} \end{cases} \tag{3.78}$$

The FSC $\hat{S}_L(k)$ of $\hat{r}_L(n)$ in step 1 are

$$\hat{S}_L(k) = \hat{S}(e^{j\omega})\Big|_{\omega=\frac{2\pi}{L}k} = 1 + \frac{5}{3}\rho\cos\frac{2\pi}{L}k, \quad k = 0, \ldots, L-1. \tag{3.79}$$

Now, assume $L = 12 > 10$, so that $K = 3$ and $P = 1$. So we have the following sets to consider in step 2:

$$\{\hat{S}_L(0), \hat{S}_L(3), \hat{S}_L(6), \hat{S}_L(9)\}, \quad \{\hat{S}_L(1), \hat{S}_L(4), \hat{S}_L(7), \hat{S}_L(10)\} \tag{3.80}$$

which are evaluated below respectively:

$$\{1 + \frac{5}{3}\rho, 1, 1 - \frac{5}{3}\rho, 1\}, \quad \{1 + \frac{5\sqrt{3}}{6}\rho, 1 - \frac{5}{6}\rho, 1 - \frac{5\sqrt{3}}{6}\rho, 1 + \frac{5}{6}\rho\}. \tag{3.81}$$

First assume $\rho > 0$. The maximum of the first set is $\hat{S}_L(0)$ and the maximum of the second set is $\hat{S}_L(1)$. Hence applying step 3 of the algorithm, we have

$$\{F_L(k),\ k = 0, \ldots, L-1\} = \{4, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4\} \tag{3.82}$$

Taking the inverse DFT of $F_L(k)$, we calculate in step 4:

$$\{f_L(n), n = 0, \ldots, N\} = \{1, \frac{1+\sqrt{3}}{3}, \frac{2}{3}, \frac{1}{3}, 0, \frac{1-\sqrt{3}}{3}\} \tag{3.83}$$

Hence the product filter $g(n) = w(n)f_L(n)$ has been found, and

$$\begin{aligned} G(z) \;=\; & \frac{1-\sqrt{3}}{18}z^5 + \frac{1}{6}z^3 + \frac{4}{9}z^2 + \frac{5(1+\sqrt{3})}{18}z + 1 \\ & + \frac{5(1+\sqrt{3})}{18}z^{-1} + \frac{4}{9}z^{-2} + \frac{1}{6}z^{-3} + \frac{1-\sqrt{3}}{18}z^{-5} \end{aligned} \tag{3.84}$$

Next consider the case $\rho < 0$. Referring to (3.81), $\hat{S}_L(6)$ in the first set and $\hat{S}_L(7)$ in the second set is maximum. Hence,

$$\{\hat{F}_L(k),\ k = 0, \ldots, L-1\} = \{0, 0, 0, 0, 0, 4, 4, 4, 0, 0, 0, 0\} \tag{3.85}$$

which is equal to $F_L(k-6)$ where $F_L(k)$ is the previous solution. Hence $\hat{f}_L(n) = (-1)^n f_L(n)$ and therefore $\hat{G}(z) = G(-z)$. By spectrally factorizing the product filter,

an optimum compaction filter is obtained. The compaction gain is

$$1 + \frac{5(1+\sqrt{3})}{9}|\rho| \simeq 1 + 1.5178|\rho| \tag{3.86}$$

Let us find the improvement we can get by optimizing the window when we fix $f_L(n)$. Since $\sigma_x^2 = r(0) = 1$, the compaction gain is the maximum eigenvalue of the 6×6 symmetric Toeplitz matrix with the first row $[1\ f_L(1)\ \rho\ 0\ 0\ 0\ 0]$ which is $1 + 1.8019 f_L(1)|\rho|$. Using $f_L(1)$ given in (3.83), the improved compaction gain is $1 + 1.6410|\rho|$. With this optimum window fixed, one can verify that $f_L(n)$ in (3.83) is still the optimum sequence.

3.4.2 Choice of the Periodicity L

The window method will produce compaction filters as long as L is a multiple of M and is greater than N. This choice of L will ensure that $f_L(n)$ is Nyquist(M). The smallest such period is $L = M\lceil N/M \rceil$ and the largest is $L = \infty$. The choice $L = M\lceil N/M \rceil$ leads to an additional symmetry in $f_L(n)$ and according to our experience, the corresponding compaction gains are not good. If we use $L = \infty$, then we get the ideal solution for $f_L(n)$: $f_L(n) = g_i(n)$. The corresponding compaction filter obtained after windowing is not optimal either. If L is chosen to be the smallest multiple of M such that $L \geq 2N$, then we obtain very good compaction gains. This choice can be compactly written as

$$L = M\lceil 2N/M \rceil$$

If $M = 2$, then this choice reduces to $L = 2N$. In Example 8, we increased L from 12 to 16 and found that the compaction gain decreased! When we used the ideal filter for $f_L(n)$ which corresponds to $L = \infty$, the compaction gain was better than that of the case $L = 16$ but worse than that of the case $L = 12$.

Example 9: Dependence on L. We have designed compaction filters using the window method for an AR(5) process whose psd is shown in Fig. 3.6. We have chosen this psd because it is multiband, and the capturing of the signal energy can be illustrated clearly. The number of channels is $M = 2$. We considered the filter orders $N = 1, 3, 5$, and $N = 31$. For each order N, we increased L from $2N$ to 100 in steps of 2. The

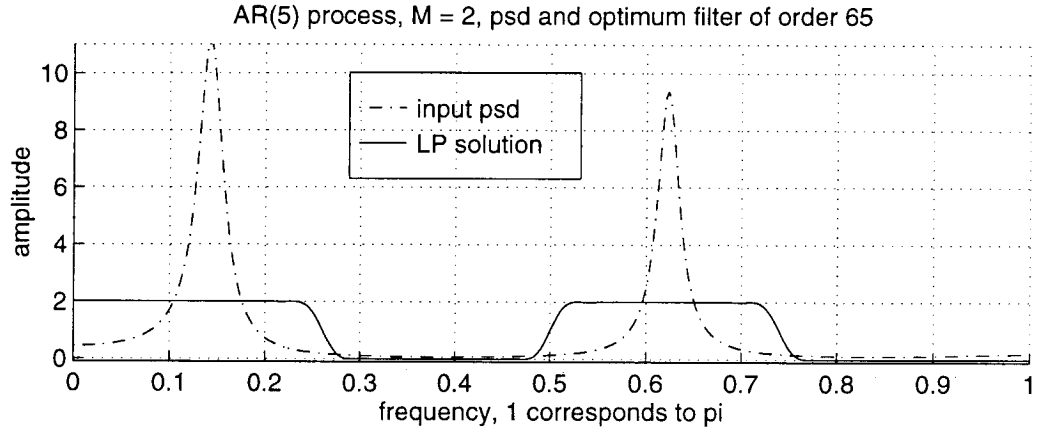


Fig. 3.6: The psd of an AR(5) process, and the magnitude square of an optimal compaction filter for $N = 65$ and $M = 2$, designed by LP. The parameter L is 512 and a triangular window is used.

resulting compaction gains are plotted in Fig. 3.7. From the plot, we see that the best

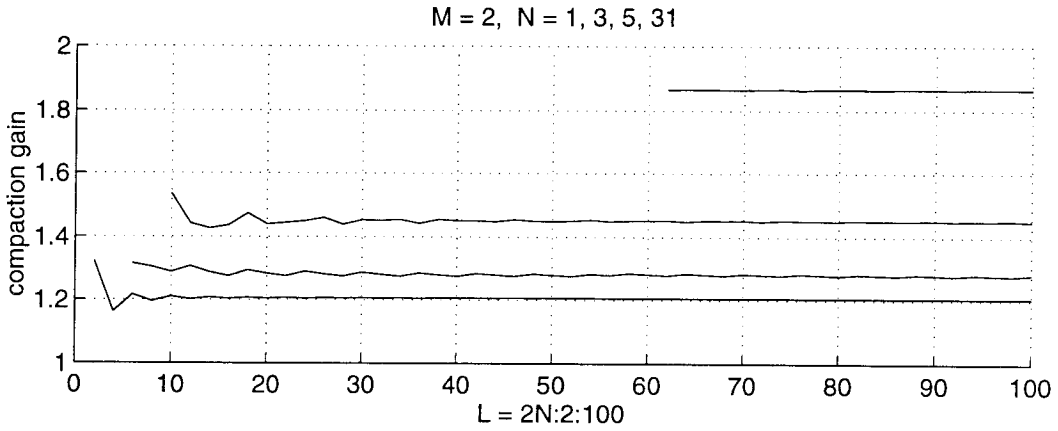


Fig. 3.7: Compaction gain vs. periodicity L.

compaction gain is for $L = 2N$.

3.5 Linear Programming Method and Multistage Designs

The use of linear programming (LP) method in compaction filter design was proposed by Moulin et al. [MAKP97]. We briefly review the method and propose some improvements. Assume that the input process $x(n)$ is real. The output variance is

$\sigma_y^2 = r(0) + 2\sum_{n=1}^{N} g(n)r(n)$. Let $\mathbf{g_d}$ and $\mathbf{r_d}$ be the vectors formed by deleting every Mth coefficients of $g(n)$ and $r(n)$ for $n = 1, \ldots N$. Then the objective can be written as $\sigma_y^2 = r(0) + 2\mathbf{r_d}^T\mathbf{g_d}$. This incorporates the Nyquist(M) condition but not the nonnegativity constraint in (3.9). Let

$$\mathbf{c_d}(\omega) \triangleq [\cos(\omega)\ \cos(2\omega)\ \ldots\ \cos((M-1)\omega)\ \cos((M+1)\omega)\ \ldots\ \cos(N\omega)]^T \quad (3.87)$$

Then $G(e^{j\omega}) = 1 + 2\mathbf{c_d}^T(\omega)\mathbf{g_d}$. Hence the problem is equivalent to:

$$\text{maximize } \mathbf{r_d}^T\mathbf{g_d} \quad \text{subject to} \quad \mathbf{c_d}^T(\omega)\mathbf{g_d} \geq -0.5,\ \forall\omega \in [0, \pi] \quad (3.88)$$

This type of problem is typically classified as semi-infinite linear programming (SIP) [MAKP97] because there are infinitely many inequality constraints on finitely many variables. By discretizing the frequency, one reduces this to a well known standard LP problem.

Drawbacks of the Technique. No matter how dense the frequency grid is, LP guarantees the nonnegativity of $G(e^{j\omega})$ only on this grid. Hence one has to modify the solution to have $G(e^{j\omega}) \geq 0,\ \forall\omega$. One can numerically determine the unit-circle zeros of $G(e^{j\omega})$ and merge the pairs of zeros that are close to each other. Yet another way is to "lift" $G(e^{j\omega})$ by increasing $g(0)$ relative to other coefficients. Since $g(0)$ has to be 1, in effect we scale $g(n)$ for $n \neq 0$ by a constant $c < 1$. This can also be considered as windowing with $w(n) = c, n \neq 0$, and $w(0) = 1$. In the next section we propose another windowing technique to modify $G(e^{j\omega})$. The advantage of this is to avoid having to locate any zeros or the minimum of $G(e^{j\omega})$. The nonnegativity of $G(e^{j\omega})$ is guaranteed by that of $W(e^{j\omega})$ as in Sec. 3.4.1. If the filter order N and the number of discrete frequencies L are small, using an optimum window performs better than the other techniques. In principal, as $L \to \infty$, the LP solution approaches the optimal solution. However, as stated in [MAKP97], there will be numerical problems if L is too high. Another drawback of LP is that the complexity is prohibitively high for high filter orders. We should note here that the window method that we proposed in Sec.

3.4 does not have this problem. The window method is very fast even with very high filter orders and the resulting filters are very close to the optimal ones.

3.5.1 Windowing of the Linear Programming Solution

Let L uniform frequencies $\{\omega_k = \frac{2\pi}{L}k,\ k = 0, \ldots, L-1\}$ be used in LP and let $g_L(n)$ be the periodical expansion of the resulting product filter. Assume that $L > 2N$. Linear programming assures that $G(e^{j\omega})$ is nonnegative at the frequencies $\{\omega_k\}$. Hence the FSC $G_L(k)$ of $g_L(n)$ are nonnegative. Now consider the product

$$g(n) = w(n)g_L(n) \tag{3.89}$$

where $w(n)$ is a symmetric window of order $K < L-N$ (length $2K+1$) with nonnegative FT (see Fig. 3.8), then from Sec. 3.4.1 we conclude that $G(e^{j\omega}) \geq 0$, $\forall\omega$. The

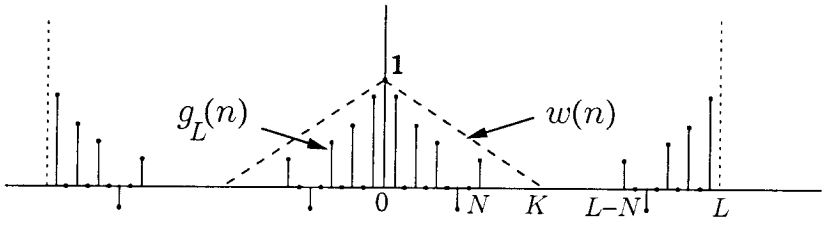


Fig. 3.8: Windowing of the linear programming solution.

Nyquist(M) property of $g(n)$ is assured by that of $g_L(n)$. In contrast to the window method, here we can have $K > N$. This is because the LP solution has already the desired order. For maximum compaction gain, the symmetric order of $w(n)$ is chosen to be maximum, namely $K = L - N - 1$. Note that when $L = 2N$, we have $g_L(N) = 2g(N)$. One can use a fixed window like a triangular window as depicted in the figure and get a satisfactory compaction gain. However, one can always optimize the window as in Sec. 3.4.1. If L is very large, optimization should be avoided as the performance loss becomes negligible. The loss can be quantified as follows: Assuming $\sigma_x^2 = 1$, when a fixed window is used the compaction gain is

$$G_w = 1 + 2\mathbf{g}^T\hat{\mathbf{r}} \tag{3.90}$$

where $\mathbf{g}$ and $\hat{\mathbf{r}}$ are the vectors formed by the sequences $g_L(n)$ and $w(n)r(n)$, $n = 0, \ldots, N$. If, for example, a triangular window of symmetric order $K = L - N - 1$ is used, we have $w(n) = 1 - \frac{n}{L-N}$, $n = 0, \ldots, N$. When the optimum window is used, the compaction gain is

$$G_o = \lambda_{max}\Big\{g_L(n)r(n)\Big\}_0^N \tag{3.91}$$

Hence the loss is

$$G_o - G_w = \lambda_{max}\Big\{g_L(n)r(n)\Big\}_0^N - 2\mathbf{g}^T\hat{\mathbf{r}} - 1 \tag{3.92}$$

As $L \to \infty$, $w(n)r(n) \to r(n)$ and $g_L(n) \to g_{opt}(n)$. Hence $G_w \to G_{opt}$. Since $G_w \leq G_o \leq G_{opt}$, we see that $G_o \to G_{opt}$ as well. Hence $G_o - G_w \to 0$ as $L \to \infty$.

Example 10. Let the input psd be as in Fig. 3.6 and let $N = 65$ and $M = 2$. In the same figure, we plot the magnitude square $|H(e^{j\omega})|^2$ of the compaction filter $H(z)$ designed by LP. The number of frequencies used in the design process was $L = 512$. We have used triangular window of symmetric order $K = L - N - 1 = 446$ and found that the resulting compaction gain is $G_w \simeq 1.8698$. If we optimize the window the compaction gain becomes $G_o \simeq 1.8744$. Hence the loss is $G_o - G_w \simeq 0.0046$. One can verify that the compaction gain of the ideal ($L = \infty$) filter is $G_{ideal} \simeq 1.8754$.

3.5.2 Multistage FIR (IFIR) Compaction Filter Design

Let $M = M_0M_1$ and consider Fig. 3.9(a). This can be redrawn as in Fig. 3.9(b). The equivalent filter is $H(z) = H_0(z)H_1(z^{M_0})$. We will first impose the Nyquist(M) condition only on $|H(e^{j\omega})|^2$. Later we will impose Nyquist conditions on individual filters that guarantee the Nyquist(M) property of $|H(e^{j\omega})|^2$. We will describe the details of how to find $H_1(z)$ for a fixed $H_0(z)$ and vice versa, in an iterative manner.

Let $G_0(e^{j\omega}) = |H_0(e^{j\omega})|^2$, $G_1(e^{j\omega}) = |H_1(e^{j\omega})|^2$, and $G(e^{j\omega}) = |H(e^{j\omega})|^2$ with impulse responses $g_0(n)$, $g_1(n)$, and $g(n)$ respectively. Denote the orders of $H_0(z)$, $H_1(z)$, and $H(z)$ by N_0, N_1, and N respectively. Hence we have $N = M_0N_1 + N_0$. Define $\mathbf{g_0} = [g_0(0)\ g_0(1)\ \ldots\ g_0(N_0)]^T$, $\mathbf{g_1} = [g_1(0)\ g_1(1)\ \ldots\ g_1(N_1)]^T$, and $\mathbf{g} = [g(0)\ g(1)\ \ldots\ g(N)]^T$.

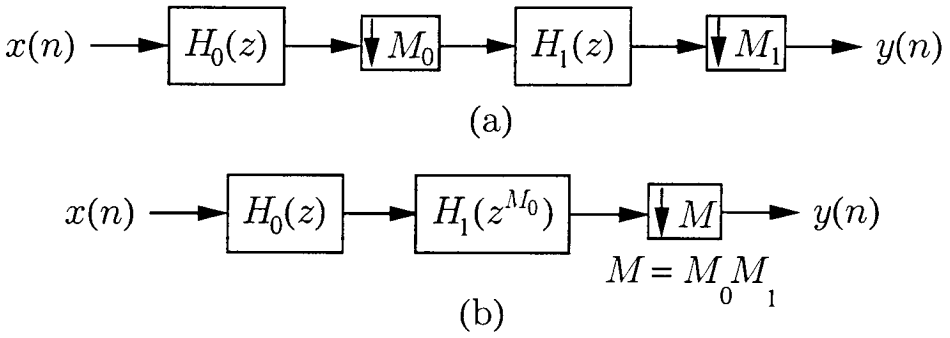


Fig. 3.9: Multistage compaction filter design. (a) Basic configuration, (b) Equivalent system.

Optimization of $H_1(z)$ for a given $H_0(z)$. We have $G(z) = G_0(z)G_1(z^{M_0})$. Let $\mathbf{G_0}$ be the $(2N+1)\times(2M_0N_1+1)$ convolution matrix formed by $g_0(n)$. Taking into account the symmetries and the fact that $G_1(z^{M_0})$ has nonzero components only for multiples of M_0, we can write $\mathbf{g} = \mathbf{A_0 g_1}$, where $\mathbf{A_0}$ is an $(N+1)\times(N_1+1)$ matrix that is obtained from $\mathbf{G_0}$. Now, the Nyquist(M) constraint requires that if we decimate $\mathbf{g}$ by M we should get $\mathbf{e_0} = [1\ 0\ \ldots\ 0]^T$. Let $\mathbf{B_0}$ denote the matrix that is obtained by taking every Mth row of $\mathbf{A_0}$. Then we should have $\mathbf{B_0 g_1} = \mathbf{e_0}$. To force the nonnegativity constraint on $G_1(e^{j\omega})$, let

$$\mathbf{c_0}(\omega) \triangleq [1\ 2\cos(\omega)\ 2\cos(2\omega)\ \ldots\ 2\cos(N_1\omega)]^T \tag{3.93}$$

Then the constraint $G_1(e^{j\omega}) \geq 0$ becomes

$$\mathbf{c_0}^T(\omega)\mathbf{g_1} \geq 0,\ \forall\omega \in [0,\pi] \tag{3.94}$$

If $\mathbf{r} = [r(0)\ 2r(1)\ \ldots\ 2r(N)]^T$, the objective is to maximize $\mathbf{r}^T\mathbf{g} = \mathbf{r}^T\mathbf{A_0 g_1}$. Hence we have reduced the problem to the following:

$$\text{maximize } \mathbf{r_0}^T\mathbf{g_1},$$
$$\text{subject to } \mathbf{B_0 g_1} = \mathbf{e_0}, \text{ and } \mathbf{c_0}^T(\omega)\mathbf{g_1} \geq 0,\ \forall\omega \in [0,\pi],$$

where $\mathbf{r_0} = \mathbf{A_0}^T\mathbf{r}$. Hence a standard linear programming algorithm can be applied, once a set of frequencies is chosen for the inequality constraint.

Optimization of $H_0(z)$ for a given $H_1(z)$. Similarly, one can reduce the problem of

finding the best $H_0(z)$ for a given $H_1(z)$ to the following linear programming problem:

$$\text{maximize } \mathbf{r_1}^T\mathbf{g_0},$$
$$\text{subject to } \mathbf{B_1 g_0} = \mathbf{e_0}, \text{ and } \mathbf{c_1}^T(\omega)\mathbf{g_0} \geq 0, \ \forall\omega \in [0,\pi],$$

where

$$\mathbf{c_1}(\omega) = [1 \ 2\cos(\omega) \ 2\cos(2\omega) \ \ldots \ 2\cos(N_0\omega)]^T \tag{3.95}$$

and $\mathbf{r_1} = \mathbf{A_1}^T\mathbf{r}$. The $(N+1)\times(N_0+1)$ matrix $\mathbf{A_1}$ is obtained from the $(2N+1)\times(2N_0+1)$ convolution matrix formed by $g_1(n)$ by taking the symmetries into account and the matrix $\mathbf{B_1}$ is obtained by taking every Mth row of $\mathbf{A_1}$.

One can iterate between the above two optimization steps until there is no significant change in the compaction gain. The initial choice of $g_0(n)$ can significantly affect the resulting compaction gain. According to our design experience, if $g_0(n)$ is chosen to be a triangular sequence, the compaction gain at the end of the iteration is very good. The filters $g_0(n)$ and $g_1(n)$ which result from the iteration should spectrally be factorized to identify $H_0(z)$ and $H_1(z)$. This step will be successful only if the solutions are such that $G_0(e^{j\omega}) \geq 0$ and $G_1(e^{j\omega}) \geq 0$ for all ω. If this is not the case, we can force it by use of windowing on $g_0(n)$ and $g_1(n)$ as described in Sec. 3.3.1 or by the "lifting" technique. If this is done then the product filter $G_0(z)G_1(z^{M_0})$ will not be exactly Nyquist(M). In the next subsection we show how to overcome this problem.

Example 11. Let us design IFIR compaction filters for the pair $(M,N) = (36,65)$, and for the input process whose psd was given in Fig. 3.6. Let $M_0 = 9$ and $M_1 = 4$, and let $N_0 = 11$ so that $N_1 = 6$. The number of frequencies used in the designs is $L = 1024$. Starting with a triangular sequence for $g_0(n)$, the algorithm converges in a few steps. We windowed the resulting solutions $g_0(n)$ and $g_1(n)$ with triangular windows of symmetric orders $L-N_0-1$ and $L-N_1-1$ respectively. The final product filter was not exactly Nyquist(M) because it was found that $g(36) \simeq -0.0018 \neq 0$. The final compaction gain was 5.1444. If we design a compaction filter of order 18 directly (i.e., not using IFIR technique), the compaction gain is 4.4225. This corresponds to a compaction filter with the same number of active multipliers, namely 19. If we design a compaction filter of order 65 directly (66 active multipliers), then the resulting

compaction gain is 7.2337.

A Particular IFIR Configuration

In Fig. 3.9, if $G_0(z)$ is Nyquist(M_0) and $G_1(z)$ is Nyquist(M_1), it can be verified that $G(z)$ given by $G_0(z)G_1(z^{M_0})$ is Nyquist(M). Now, let us fix $H_0(z)$ to be a valid compaction filter for the pair (N_0, M_0). Referring to Fig. 3.10(a), the best $H_1(z)$ is the optimum compaction filter for (N_1, M_1), and for the input $x_0(n)$ which has the psd $S_{x_0x_0}(z) = \left(G_0(z)S_{xx}(z)\right)\Big|_{\downarrow M_0}$. Similarly, if $H_1(z)$ is a fixed compaction filter for

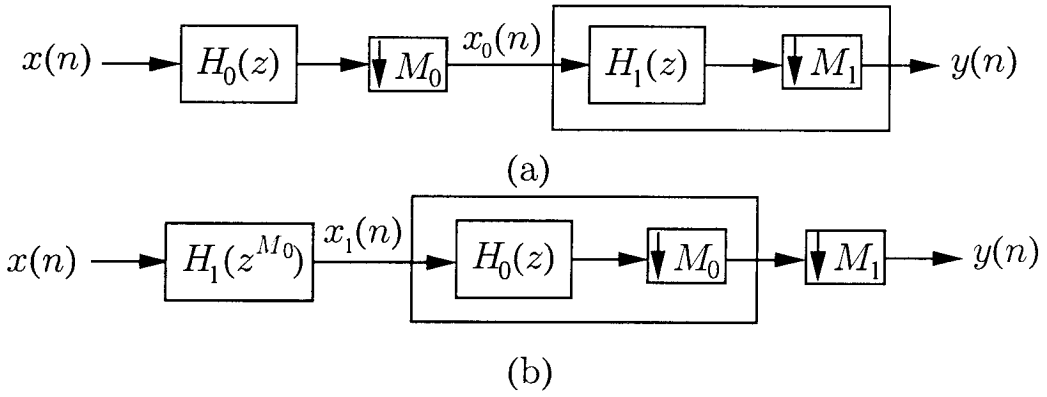


Fig. 3.10: Special IFIR design configuration.

the pair (N_1, M_1), then we can redraw the configuration as in Fig. 3.10(b). The best $H_0(z)$ is the optimum compaction filter for (N_0, M_0), and for the input $x_1(n)$ which has the psd $S_{x_1x_1}(z) = G_1(z^{M_0})S_{xx}(z)$. One can design the compaction filters $H_0(z)$ and $H_1(z)$ iteratively using any of the known techniques. Hence, one can use the linear programming technique as well as any other technique like the noniterative methods to be mentioned in the next section.

Example 12. Let the setup be the same as in the previous example. We have designed the compaction filters $H_0(z)$ and $H_1(z)$ iteratively using the standard linear programming procedure as in Example 10. We have started with $H_1(z) = 1$. The first compaction filter $H_0(z)$ is therefore the optimal compaction filter for the pair $(M_0, N_0) = (9, 11)$ for the original autocorrelation sequence. We have windowed the final product filters as we did in Example 10 to guarantee the nonnegativity. The resulting overall compaction gain is 4.9432. This is slightly smaller than the overall

compaction gain 5.1444 in Example 10. However, the resulting overall filter here is exactly Nyquist(M) unlike the case of Example 11.

3.6 Comparison of Methods

3.6.1 Connection Between the Linear Programming Method and the Window Method

In both the LP and window methods, we use windows to assure the nonnegativity of $G(e^{j\omega})$. Consider the equations (3.89) and (3.66). When L is a multiple of M, a periodic sequence $g_L(n)$ in the linear programming method, and a periodic sequence $f_L(n)$ in the window method are found such that they are Nyquist(M) and their FSC are all nonnegative. For $L > 2N$, the two problems are not the same because $g_L(n)$ is order constrained while $f_L(n)$ is not. If, however, $L = 2N$, then the two problems are exactly the same! If windowing is done in the same way in both methods, then we see that the resulting compaction gains should be the same. Hence, one can view the window method as an efficient and noniterative technique to solve an LP problem when $L = 2N$. If L is increased, we saw that the window method does not necessarily yield better gains whereas this is the case for the LP method provided the window order is increased as well. However, optimization of the window in LP becomes costly as the order increases. If one uses a fixed triangular window (with highest possible order) in LP, and if the windows are optimized in the window method, then window method is very close and sometimes superior to LP as we demonstrate in the following example.

Example 13: Comparison of linear programming and window methods. Let the input psd be as in Fig. 3.6. In Fig. 3.11(a) the compaction gains of both the LP and the window method versus the filter order are plotted for $M = 2$. The number of frequencies used in LP is $L = 512$ while the periodicity used in the window method is $L = 2N$. The windows used in LP are triangular windows with symmetric order $L - N - 1$. In the window method, the autocorrelation sequence is first windowed by a triangular window of symmetric order N to find $f_L(n)$ and then the window is re-

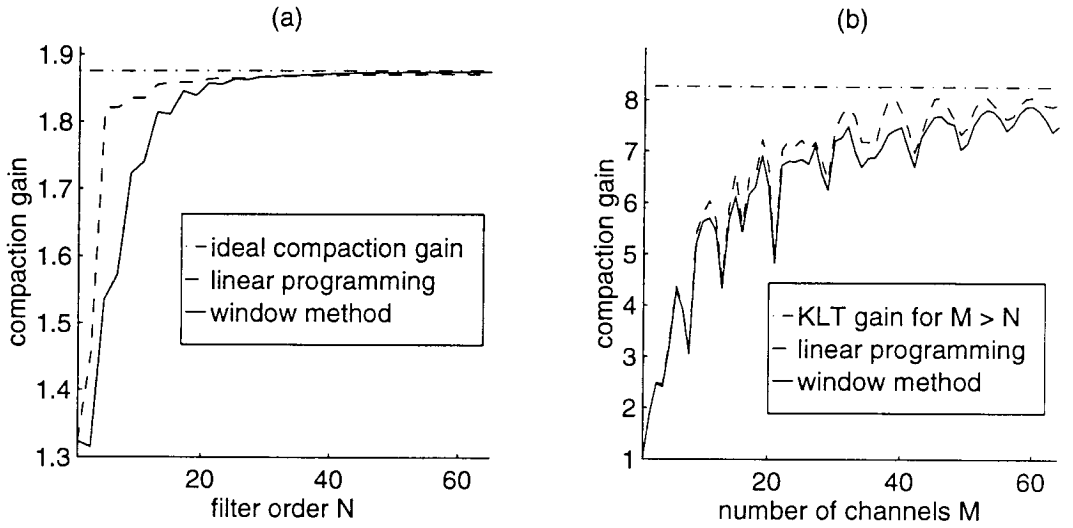


Fig. 3.11: Comparison of the window and linear programming methods. The input power spectrum is as shown in Fig. 3.6. (a) Compaction gain versus N, for $M = 2$, (b) Compaction gain versus M, for $N = 65$.

optimized. From the figure we observe that if the order is high, one has slightly better compaction gains using the window method. This implies that, if one optimizes the window, there is no need to use large number of frequencies in LP! More importantly, there is no need to use LP for high filter orders. However, it should be emphasized that if the windows are optimized in LP, one can get slightly better compaction gains than the window method. In Fig. 3.11(b), we show the plots of the compaction gains of the two methods for various values of M for a fixed filter order of 65. We observe that the window method performs very close to LP especially for low values of M. We show the upper bounds on compaction gains in both plots. The upper bound in the first plot is achieved by an ideal compaction filter and that in the second plot is achieved by a maximal eigenfilter as discussed in Sec. 3.3.1.

Example 14. Let the input be AR(1) as in Example 1. For $N = 3$ and $M = 2$, we have designed compaction filters using the window and LP methods. We present in Table 3.1 the resulting filter coefficients and the corresponding compaction gains for $\rho = 0.1, 0.5$, and 0.9. Also presented in the same table are the analytically optimum coefficients (3.46), and the corresponding compaction gains (3.48). We see in this case that the compaction gain of the window method is not too far from the optimal one and

slightly worse than that of LP even for such a small order. The discrepancy between the window and LP compaction gains is maximum when $\rho = 0.5$.

3.7 Concluding Remarks

We have presented new techniques for the design of optimum FIR compaction filters. First we have proposed an analytical method in the two-channel case. The technique is applicable for a rather restricted but practically important class of signals. The method involves Levinson recursion and two spectral factorizations of half the filter order. As examples we have produced analytical expressions for the compaction filter coefficients for AR(1) and MA(1) processes. Next we have proposed a method called the **window method**. It is applicable for any given spectra and for any given number of channels. It is very efficient since it is noniterative and involves only comparison of some DFT coefficients and windowing. We have given its relation to the LP method. As the filter order becomes higher, the computational complexity of the LP method grows rapidly. The window method on the other hand is very fast even when the filter orders are very high. Furthermore, the suboptimality of the window method diminishes as the filter order increases. Finally we have presented multistage design techniques that enable the design of a compaction filter with a given order in multiple stages each involving the design of a smaller order compaction filter.

We believe that the methods we presented in this chapter can be incorporated in the design of optimal FIR orthonormal uniform and nonuniform subband coders. In the two channel case, the optimum compaction filter already determines the optimum filter bank. Hence the algorithms in this chapter can readily be used in applications like wavelet-based image coding. In particular, it would be interesting to investigate the performance of our filters in zero-tree coding and wavelet-package coding. For such applications, we expect that the analytical method of Sec. 3.3.2 will be quite useful. In the $M-$channel case, we mentioned one method [MM98] that efficiently finds the rest of the filter bank optimally if the first filter is given. In speech and audio coding applications, $M-$channel uniform filter banks are commonly used and

the filters have high orders. We expect that the window method of Sec. 3.4 will be very useful for such applications. Needless to say, there are many other important applications of compaction filters, some of which are mentioned in the last paragraph of the introduction of the chapter. Hence our design algorithms can directly be used in such applications as well. All the algorithms described in this chapter can be found at our webpage [htt].

APPENDIX

Proof of nonnegativity. We will show that $G(e^{j\omega})$ obtained by the procedure in Sec. 3.3 is necessarily nonnegative in the region $[\pi/2, \pi]$. The Nyquist(2) property of $G(e^{j\omega})$ implies $G'(e^{j\omega}) = G'(e^{j(\pi-\omega)})$. We therefore have

$$G'(e^{j\omega_k/2}) = G'(e^{j(\pi-\omega_k/2)}) = 0, \quad k = 0, \ldots, (N-3)/4 \tag{3.96}$$

Now, by the mean value theorem in calculus, we also have

$$G'(e^{j\hat{\omega}_k/2}) = G'(e^{j(\pi-\hat{\omega}_k/2)}) = 0, \quad k = 0, \ldots, (N-7)/4 \tag{3.97}$$

for some $\hat{\omega}_k \in (\omega_k, \omega_{k+1})$. Notice that since ω_k's are all distinct and lie in the open region $(0, \pi)$, all of the above zeros are distinct. The total number of such zeros is therefore $N-1$. Since $G(e^{j\omega})$ is a cosine polynomial of order N, $G'(e^{j\omega})$ is a sine polynomial of order N and therefore it can be written in the form

$$G'(e^{j\omega}) = \sin\omega \; T(\cos\omega) \tag{3.98}$$

where $T(x)$ is a polynomial of order $N-1$. Excluding the zeros at 0 and π, the total number of zeros $G'(e^{j\omega})$ can have in $[0, \pi]$ is $N-1$. Hence $G'(e^{j\omega})$ cannot have any other zero on the unit-circle. If $G(z)$ has a zero at $\pi - \omega_k/2$ with multiplicity greater than 2, then $G'(e^{j\omega})$ has at least double zero at that frequency implying that the total number of its zeros is more than $N-1$ which is a contradiction. If $G(z)$ has a single zero in the region $(\pi/2, \pi)$ which is different from all ω_k's, then, by applying the mean value theorem once more, $G'(e^{j\omega})$ has to have another zero which is again a contradiction.

Hence we have proved that $G(e^{j\omega})$ has double zeros at

$$\pi - \omega_k/2, \quad k = 0, \ldots, (N-3)/4 \tag{3.99}$$

and that it does not have any other unit-circle zeros in $[\pi/2, \pi]$. This in particular implies $G(e^{j\omega}) \geq 0$ for $\omega \in [\pi/2, \pi]$. The proof for the case of even $\frac{N-1}{2}$ is similar; the details are omitted. ■

Chapter 4

Lattice Quantization and Vector Dithering

Lattice vector quantizers have recently become attractive because they are simple to implement and, in most cases, they constitute good alternatives to the computationally more complex vector quantization algorithms like the LBG and ECVQ [GG92, CLG89]. The geometric regularity of lattices allow very fast quantization algorithms, and there are already efficient algorithms for several well-known lattice structures [CS82, SGR84, GS88, JG93, JG95].

Dithering was first applied by Roberts [Rob62] to image coding. It was seen that by adding an independent random variable called dither before the quantization and subtracting after it, the perceptual quality of the image improves substantially. After that pioneering idea, there has been considerable work on the theory and applications of dithering. Dithered quantizers were theoretically analyzed by Schuchman [Sch64] using the so called characteristic function method which uses the Fourier transform of the input probability density function (pdf). An analysis of the undithered uniform quantization was provided by Sripad and Snyder [SS77], using a similar style. More recently, Lipshitz et al. [LWV92] published an excellent survey on quantization and dither. Gray and Stockham [GS93] gave new insightful proofs for the cases of subtractive and nonsubtractive dithering.

In this chapter we consider the idea of dithering in lattice quantization. The idea has

already been introduced by Ziv [Ziv85] as a means of universal quantization. Interesting results on the rate distortion efficiency of dithered lattice quantizers have already been obtained by Zamir and Feder [ZF92, ZF94, ZF95, ZF96], and by Linder and Zeger [LZ94]. In this chapter our major concern is the analysis of the lattice quantization error for dithered and undithered cases. The only overlap between our work and the literature that we are aware of is Theorem 5. This was also reported by Zamir and Feder as a small part of their recent paper [ZF95]. Even in our work, this result, independently found by us, is only a minor ingredient.

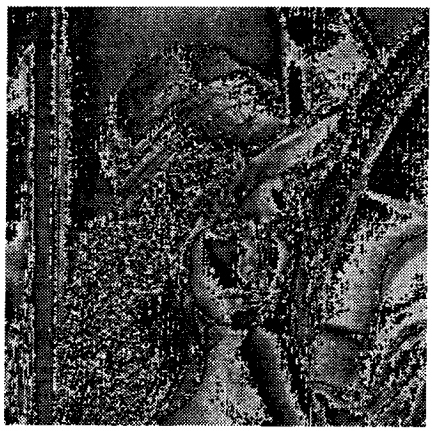
In Sec. 4.1, we review some preliminaries and definitions pertaining to lattice quantization. In Sec. 4.2, we provide exact analysis of the lattice quantization system. This can be regarded as a multidimensional extension of the work in [SS77]. The main tool, accordingly, is again Fourier series, but this time multidimensional. Since lattices are uniform structures, there is inherent periodicity in the error statistics, which motivates the use of multidimensional Fourier series. However, unlike in the one-dimensional case, the choice of lattice is no longer unique and there exist optimum lattices in the sense that they minimize the familiar dimensionless second moment [LZ94]. After giving the exact relationships between input and error probability densities, we consider dithered, or so called randomized lattice quantization schemes. As in one-dimensional case [LWV92], we investigate the possibility of rendering error statistics independent from the input. Sec. 4.3 covers subtractive dithering where an appropriate random vector is added before the quantizer and subtracted after it. In Sec. 4.4 nonsubtractive dithering is examined. Sec. 4.5 is devoted to finding optimum linear time invariant pre- and post-filters to be used in conjunction with dithered lattice quantizers.

Demonstration of the Perceptive Advantages of Vector-dithering in Image Coding

For motivational purposes, we show in Fig. 4.1 a demonstration of the improvement of perceptual quality in image compression achieved by the use of vector-dithered lattice quantization. Fig. 4.1(a) shows the original 8 bit/pixel, 512×512 image of Lenna. Fig. 4.1(b) shows the output of a lattice quantizer with dimension 24. Vectors are

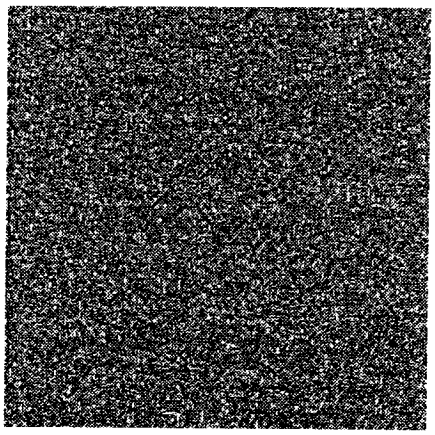
(a)

(b)



(c)

(d)



(e)

Fig. 4.1: Demonstration of the perceptual advantages of dithered lattice quantizers. (a) Original image of Lenna, 512x512, 8 bits/pixel. (b) Output of lattice quantization with dimension 24, bit rate = .4 bits/pixel. (c) Error of the lattice quantization. (d) Output of the same lattice quantization with subtractive dithering, bit rate is about the same. (e) Error of the dithered lattice quantization.

formed by taking 4×6 blocks. The bit rate is about 0.4 bits/pixel. Fig. 4.1(c) shows the quantization error. Fig. 4.1(d) shows the output of the same lattice quantizer but with subtractive dithering. The bit rate is about the same. The corresponding quantization error is shown in Fig. 4.1(e). It is clear that the lattice quantization error in Fig. 4.1(c) is highly correlated with the input while the dithered quantization error in Fig. 4.1(e) seems completely uncorrelated with the input and uniform. The output of dithered lattice quantizer is perceptually more pleasant than that of undithered one.

Summary of the Main Results of the Chapter

1. In Sec. 4.2 we provide the necessary and sufficient condition for the quantization error of an undithered lattice quantizer to be uniform in its quantization basic cell. This is the so-called Nyquist-$\mathbf{V}$ condition, where $\mathbf{V}$ is the lattice generator matrix. We provide examples and general classes of random vectors that satisfy this condition (Sec. 4.2.1). We then examine the error statistics when the input is arbitrary (Sec. 4.2.2).

2. We next consider subtractive vector dithering, and establish the necessary and sufficient condition for the quantization error to be statistically independent of the input, and be uniform in the quantization basic cell. A comparison of the dimensionless second moment of lattice quantizers [LZ94] is then given. A necessary condition for a lattice quantizer to have minimum dimensionless second moment (among all lattice quantizers of the same dimension) is established (Theorem 5).

3. For nonsubtractive vector dithering, first and second order moments of the quantization error conditioned on the input vector are derived (Sec. 4.4). Necessary and sufficient conditions for these moments to be independent of the input are provided. Examples of nonsubtractive dither vectors satisfying the moment independence conditions are given, and the dither that produces the minimum error for a given lattice is distinguished (Theorem 7).

4. In Sec. 4.5 we consider the use of a linear pre-filter prior to the lattice quantization of a wide sense stationary (WSS) vector random process $\mathbf{x}(n)$. Under the

assumption that the lattice quantizer satisfies certain mild conditions, we will derive an expression for the best choice of pre-filter, as a function of the power spectral density matrix of the input process. We will also clarify the similarity and differences between this problem and the problem of designing optimal biorthogonal subband coders.

The results of this chapter have been published in a journal paper [KV96b] and some of the results are presented at a conference [KV95].

4.1 Preliminaries and Definitions

Let R^D and Z^D denote the D-dimensional Euclidean space of real numbers and the D-dimensional space of integers respectively. Let $\mathbf{V} = [\,\mathbf{v}_1 \quad \mathbf{v}_2 \quad \ldots \quad \mathbf{v}_D\,]$ be a nondegenerate lattice base in R^D. The lattice is the set of vectors defined as

$$\mathcal{L}(\mathbf{V}) = \{\mathbf{x} : \mathbf{x} = \mathbf{V}\mathbf{n}, \quad \mathbf{n} \in Z^D\} \tag{4.1}$$

Fig. 4.2 shows an example of a lattice in two dimensions. In lattice quantization, the

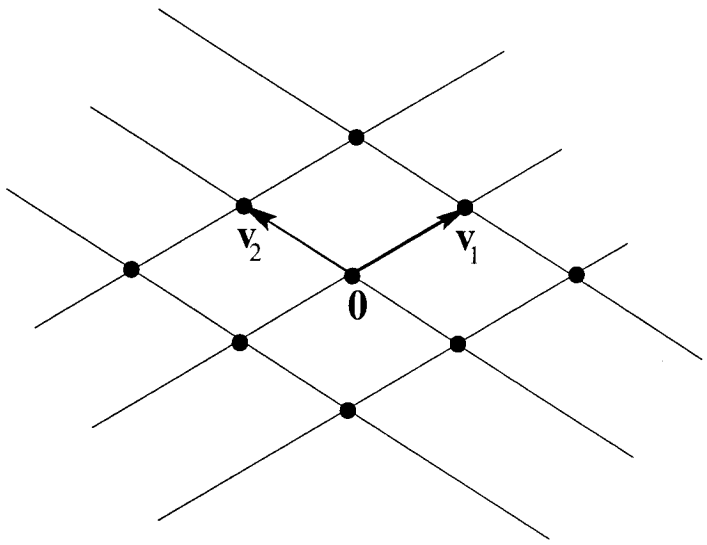


Fig. 4.2: Lattice example in $2-D$. The heavy dots are the points on the lattice.

codewords are the lattice points. The partition of the space for decision regions can

be done in many ways. This partitioning can be uniform, i.e., each codeword may have the same quantization cell called a *basic cell* (defined below) [Cox69]. From the necessary conditions for distortion-minimal quantizers [GG92], the quantization cell should be the so called Voronoi region [CS88] which is defined below. The resulting uniform partition is also known as the nearest-neighbor partition. Note that, from the same necessary conditions, codewords should be the centroids of the quantization cells with respect to the given distortion measure and the input probability density function. However, as in the uniform scalar quantization, one chooses lattice points as reproduction points avoiding the knowledge of probability density function.

If overflow is avoided at all times, then we have a periodic structure for the quantization error and the tools of the following analysis are applicable. In this chapter, overflow is always assumed to be avoided. If one uses entropy coding [CS88] after the quantization, or if the given density has finite support, the resulting bit rate will be finite and by scaling the lattice one can tradeoff bit rate against distortion.

Definition 1. A *basic cell* of a lattice $\mathcal{L}(\mathbf{V})$: Let $\mathcal{P}$ be a region in R^D such that any $\mathbf{x} \in R^D$ can be written as $\mathbf{x} = \mathbf{x}_0 + \mathbf{V}\mathbf{n}$ for a unique $\mathbf{x}_0 \in \mathcal{P}$ and $\mathbf{n} \in Z^D$. Then $\mathcal{P}$ is called a basic cell of the lattice $\mathcal{L}(\mathbf{V})$. It is also said to generate a tiling of R^D with respect to $\mathbf{V}$. $\diamondsuit$

This definition does not imply that a basic cell is convex. In fact, one can partition a convex basic cell into subregions, and then translate each of these subregions by some distinct lattice vectors. The resulting nonconvex region is another basic cell.

Definition 2. The *Voronoi region* of a lattice point $\mathbf{x}_0 \in \mathcal{L}(\mathbf{V})$ is the set of points that are nearer (with respect to Euclidean distance) to that point than to any other lattice point. That is,

$$VOR(\mathbf{x}_0) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\| \leq \|\mathbf{x} - \mathbf{V}\mathbf{n}\|, \quad \forall \mathbf{n} \in Z^D\} \tag{4.2}$$

$\diamondsuit$

The Voronoi region of the lattice point $\mathbf{0}$, $VOR(\mathbf{0})$, will be denoted by $VOR(\mathbf{V})$ for convenience. Fig. 4.3 shows the $VOR(\mathbf{V})$ of the lattice given in Fig. 4.2.

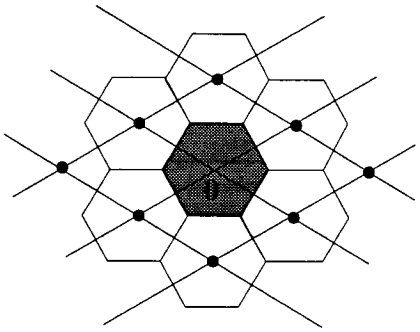


Fig. 4.3: Voronoi regions for the lattice in Fig. 4.2. The shaded region is $VOR(\mathbf{V})$.

The Euclidean distance, used in the definition, leads to the mean square error as a distortion measure. In this chapter, our interest will be only in the mean square error.

Definition 3. The *Symmetric Parallelpiped* of a lattice point $\mathbf{x}_0 \in \mathcal{L}(\mathbf{V})$ is defined as [Vai93]:

$$SPD(\mathbf{x}_0) = \{\mathbf{x} : \mathbf{x} = \mathbf{x}_0 + \mathbf{V}\mathbf{u}, \quad \forall \mathbf{u} \in [-\frac{1}{2}, \frac{1}{2})^D\} \tag{4.3}$$

$\diamondsuit$

We will denote the Symmetric Parallelpiped region of the lattice point $\mathbf{0}$, $SPD(\mathbf{0})$, by $SPD(\mathbf{V})$. Fig. 4.4 shows the $SPD(\mathbf{V})$ of the lattice given in Fig. 4.2.

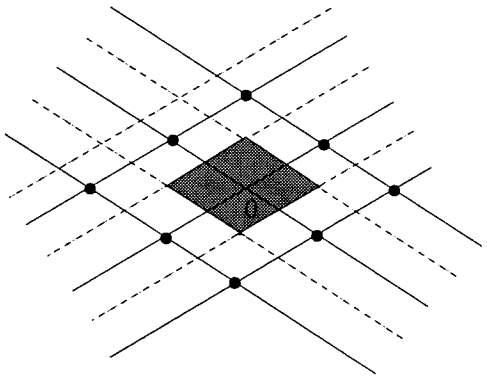


Fig. 4.4: SPD regions for the lattice in Fig. 4.2. The shaded region is $SPD(\mathbf{V})$.

It can be verified that both $VOR(\mathbf{V})$ and $SPD(\mathbf{V})$ are basic cells of the lattice $\mathcal{L}(\mathbf{V})$ as long as some modifications are done to the boundary points in order to satisfy the uniqueness requirement in the definition of a basic cell. Furthermore, they are symmetric with respect to the origin.

Definition 4. A *lattice quantizer* $Q(\mathcal{P}_0, \mathbf{V})$ with the lattice $\mathcal{L}(\mathbf{V})$ and the quantization

basic cell $\mathcal{P}_0$ is a nonlinear mapping from R^D to $\mathcal{L}(\mathbf{V})$ as given by the relation:

$$Q(\mathbf{x}) = \mathbf{V}\mathbf{n} \tag{4.4}$$

where $\mathbf{n}$ is the unique vector satisfying

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{V}\mathbf{n}, \quad \mathbf{x}_0 \in \mathcal{P}_0. \tag{4.5}$$

We will denote the quantizer with $\mathcal{P}_0 = VOR(\mathbf{V})$ by $Q(VOR, \mathbf{V})$. ♢

Definition 5. The *second moment matrix* of a basic cell $\mathcal{P}$ of a lattice $\mathcal{L}(\mathbf{V})$, denoted by $\mathbf{G}_D(\mathcal{P}, \mathbf{V})$, is the second moment of a random vector that is uniformly distributed in $\mathcal{P}$. That is,

$$\mathbf{G}_D(\mathcal{P}, \mathbf{V}) = \frac{1}{|det\mathbf{V}|} \int_{\mathcal{P}} \mathbf{e}\mathbf{e}^T d\mathbf{e}. \tag{4.6}$$

If $\mathcal{P} = VOR(\mathbf{V})$, we will denote the second moment matrix by $\mathbf{G}_D(VOR, \mathbf{V})$. ♢

Definition 6. The characteristic function of a random vector with pdf $f_{\mathbf{X}}(\mathbf{x})$ is defined to be:

$$\Phi_{\mathbf{X}}(\mathbf{\Omega}) = E[e^{j\mathbf{\Omega}^T\mathbf{x}}] = \int f_{\mathbf{X}}(\mathbf{x}) e^{j\mathbf{\Omega}^T\mathbf{x}} d\mathbf{x}. \tag{4.7}$$

♢

Next we will define a Nyquist-$\mathbf{V}$ vector. We say a function $f(\mathbf{\Omega})$ is Nyquist-$\mathbf{A}$ if $f(\mathbf{An}) = c\delta(\mathbf{n}), \forall \mathbf{n} \in Z^D$, where c is a constant and $\delta(\mathbf{n})$ is Dirac delta function, which is 1 when $\mathbf{n} = \mathbf{0}$, and 0 otherwise.

Definition 7. A *Nyquist-*$\mathbf{V}$ *random vector* is a vector whose characteristic function is Nyquist-$\mathbf{U}$, that is $\Phi_{\mathbf{X}}(\mathbf{Un}) = \delta(\mathbf{n})$, where $\mathbf{U}$ is the generating matrix of the reciprocal lattice:

$$\mathbf{U} = 2\pi\mathbf{V}^{-T} = [\,\mathbf{u}_1 \quad \mathbf{u}_2 \quad \ldots \quad \mathbf{u}_D\,] \tag{4.8}$$

♢

Note that, by definition $\Phi_{\mathbf{X}}(\mathbf{0}) = 1$. The space domain equivalent of the Nyquist condition is:

$$\sum_{\mathbf{n}} f_{\mathbf{X}}(\mathbf{x} + \mathbf{V}\mathbf{n}) = \frac{1}{|det\mathbf{V}|} \tag{4.9}$$

Examples of Nyquist-$\mathbf{V}$ random vectors are given in Sec. 4.2. Whenever the matrix $\mathbf{V}$ is clear from the context, we will just say Nyquist for both random vectors and their characteristic functions.

4.2 Quantization Analysis

Define the error vector of a lattice quantizer, $Q(\mathcal{P}_0, \mathbf{V})$, as $\mathbf{e} = \mathbf{x} - Q(\mathbf{x})$. From the definition of the lattice quantizer, this error necessarily lies in $\mathcal{P}_0$. Each error vector $\mathbf{e} \in \mathcal{P}_0$ is produced by infinitely many input vectors of the form $\mathbf{e} + \mathbf{Vn}, \quad \mathbf{n} \in Z^D$ (see Fig. 4.5). Hence, the probability density function of error is:

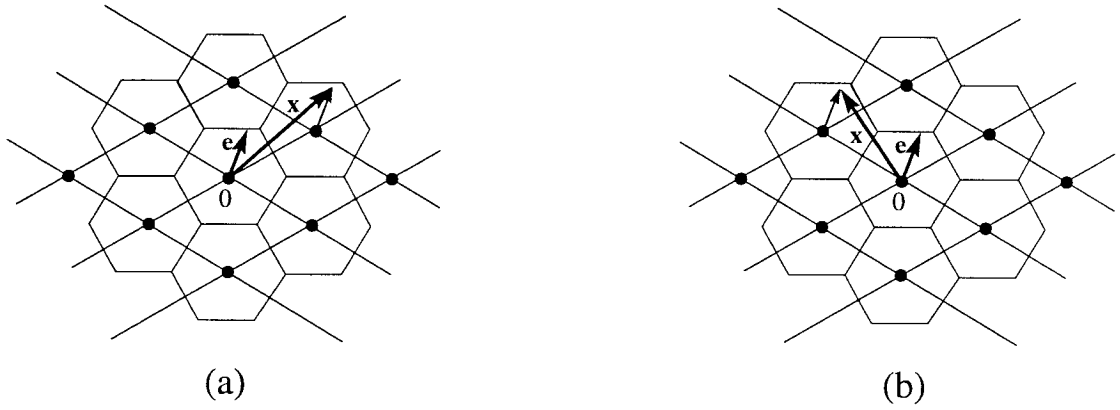


Fig. 4.5: Error in lattice quantization. (a) An error vector $\mathbf{e}$, and an input vector $\mathbf{x}$ that produces it. (b) A different input vector producing the same error vector.

$$f_{\mathbf{E}}(\mathbf{e}) = \begin{cases} \sum_{\mathbf{n}} f_{\mathbf{X}}(\mathbf{e} + \mathbf{Vn}), & \mathbf{e} \in \mathcal{P}_0; \\ 0 & \text{elsewhere} \end{cases} \tag{4.10}$$

One can find the Fourier series expansion, $\tilde{f}_{\mathbf{E}}(\mathbf{e})$, of $f_{\mathbf{E}}(\mathbf{e})$ with respect to the lattice generator matrix $\mathbf{V}$ and express it in the following form:

$$\tilde{f}_{\mathbf{E}}(\mathbf{e}) = \frac{1}{|det\mathbf{V}|} \sum_{\mathbf{n}} \Phi_{\mathbf{X}}(\mathbf{Un}) e^{-j\mathbf{e}^T\mathbf{Un}} \tag{4.11}$$

The restriction of $\tilde{f}_{\mathbf{E}}(\mathbf{e})$ to the basic cell $\mathcal{P}_0$ is $f_{\mathbf{E}}(\mathbf{e})$. For a brief summary of the relation between multidimensional Fourier series and Fourier transform, see Appendix

A. For further details of multidimensional Fourier series representation, the reader is referred to [DM84].

Theorem 1. *The quantization error of a lattice quantizer* $Q(\mathcal{P}_0, \mathbf{V})$ *is uniform in* $\mathcal{P}_0$, *that is*

$$f_{\mathbf{E}}(\mathbf{e}) = \begin{cases} \frac{1}{|det\mathbf{V}|}, & \mathbf{e} \in \mathcal{P}_0; \\ 0 & elsewhere \end{cases} \tag{4.12}$$

if and only if the input vector $\mathbf{x}$ *is Nyquist-*$\mathbf{V}$, *that is* $\Phi_{\mathbf{X}}(\mathbf{Un}) = \delta(\mathbf{n})$. ◇

Proof. From the properties of Fourier series, we know that $\tilde{f}_{\mathbf{E}}(\mathbf{e})$ in (4.11) is a constant for all $\mathbf{e}$ if and only if the Fourier series coefficients $\Phi_{\mathbf{X}}(\mathbf{Un}) = 0, \forall \mathbf{n} \neq \mathbf{0}$. Thus $f_{\mathbf{E}}(\mathbf{e})$, the restriction of $\tilde{f}_{\mathbf{E}}(\mathbf{e})$ to $\mathcal{P}_0$, is constant in $\mathcal{P}_0$ if and only if $\Phi_{\mathbf{X}}(\mathbf{Un}) = \delta(\mathbf{n})$. ■

If $\mathcal{P}_0 = VOR(\mathbf{V})$ and if the condition of the theorem is satisfied, then $E[\mathbf{e}] = \mathbf{0}$ because $VOR(\mathbf{V})$ is symmetric with respect to the origin, and $E[\mathbf{ee}^T] = \mathbf{G}_D(VOR, \mathbf{V})$, where $\mathbf{G}_D(VOR, \mathbf{V})$ is defined as in (4.6).

4.2.1 Nyquist-V Random Vectors

The next theorem shows some general classes of Nyquist-$\mathbf{V}$ vectors:

Theorem 2. *The following random vectors* $\mathbf{x}$ *are Nyquist-*$\mathbf{V}$ *and therefore have uniform quantization errors in the quantization basic cell* $\mathcal{P}_0$ *when quantized by a lattice quantizer* $Q(\mathcal{P}_0, \mathbf{V})$:

1. $\mathbf{x}$ *is uniform in any basic cell* $\mathcal{P}$ *of the lattice* $\mathcal{L}(\mathbf{V})$.

2. $\mathbf{x}$ *is piecewise uniform in an arbitrary union of nonoverlapping basic cells of* $\mathcal{L}(\mathbf{V})$, *that is*

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} c_i, & \mathbf{x} \in \mathcal{P}_i; \\ 0 & elsewhere \end{cases} \tag{4.13}$$

   where c_i*'s are positive and* $\sum_i c_i = \frac{1}{|det\mathbf{V}|}$.

3. $\mathbf{x}$ *is a sum of several independent random vectors, one of which is Nyquist-*$\mathbf{V}$.

$\diamondsuit$

Proof.

1. Let $Q(\mathcal{P}, \mathbf{V})$ be a lattice quantizer with the basic cell $\mathcal{P}_0 = \mathcal{P}$. Then, $Q(\mathbf{x}) = \mathbf{0}$, and therefore $\mathbf{e} = \mathbf{x}$. Hence $\mathbf{e}$ is uniform in $\mathcal{P}_0$. By Theorem 1, $\mathbf{x}$ is Nyquist-$\mathbf{V}$ and therefore it has a uniform quantization error in $\mathcal{P}_0$ even if it is quantized with a lattice quantizer $Q(\mathcal{P}_0, \mathbf{V})$ with $\mathcal{P}_0 \neq \mathcal{P}$.

2. Writing the characteristic function explicitly, we have:

$$\begin{aligned} \Phi_{\mathbf{X}}(\mathbf{Un}) &= \int f_{\mathbf{X}}(\mathbf{x}) e^{j\mathbf{x}^T\mathbf{Un}} d\mathbf{x} \\ &= \sum_i \int_{\mathcal{P}_i} c_i e^{j\mathbf{x}^T\mathbf{Un}} d\mathbf{x} \quad \text{(by nonoverlapping assumption)} \\ &= |det\mathbf{V}| \sum_i c_i \delta(\mathbf{n}) \quad \text{(from part 1)} \\ &= \delta(\mathbf{n}). \end{aligned} \tag{4.14}$$

3. Let $\mathbf{x} = \mathbf{v} + \mathbf{z}$, where $\mathbf{v}$ and $\mathbf{z}$ are independent. Then, $\Phi_{\mathbf{X}}(\mathbf{\Omega}) = \Phi_{\mathbf{V}}(\mathbf{\Omega})\Phi_{\mathbf{Z}}(\mathbf{\Omega})$. Therefore,

$$\Phi_{\mathbf{X}}(\mathbf{Un}) = \Phi_{\mathbf{V}}(\mathbf{Un})\Phi_{\mathbf{Z}}(\mathbf{Un}) = \delta(\mathbf{n}) \quad \text{if} \quad \Phi_{\mathbf{V}}(\mathbf{Un}) = \delta(\mathbf{n}) \quad \text{or} \quad \Phi_{\mathbf{Z}}(\mathbf{Un}) = \delta(\mathbf{n}). \tag{4.15}$$

Hence, if one of $\mathbf{v}$ or $\mathbf{z}$ is Nyquist-$\mathbf{V}$, then the sum is Nyquist-$\mathbf{V}$ as well. The extension to arbitrary number of independent random vectors is obvious.

■

Example 1. If $\mathbf{x}$ is uniform in $SPD(\mathbf{V})$ or $VOR(\mathbf{V})$, then it is Nyquist-$\mathbf{V}$ because both $SPD(\mathbf{V})$ and $VOR(\mathbf{V})$ are basic cells of the lattice $\mathcal{L}(\mathbf{V})$.

The importance of Theorem 1 rests on the fact that we can make any given input vector satisfy the Nyquist condition by applying dither prior to quantization (Sec. 4.3).

If the dither is Nyquist-$\mathbf{V}$ and independent of the input (which is quite easy to manage as we will see), then from Theorem 2, part 3 the dithered random vector is Nyquist-$\mathbf{V}$ as well.

4.2.2 Error Statistics When the Input is Arbitrary

What if the input vector is not Nyquist-$\mathbf{V}$ and we do not want to manipulate it by a dither? In that case, we have the following theorem that states the expected value of any function of the error vector $\mathbf{e}$:

Theorem 3. *Let* $\mathbf{e}$ *be the error vector of a lattice quantizer* $Q(\mathcal{P}_0, \mathbf{V})$. *Let* $g(\mathbf{e})$ *be an arbitrary function of* $\mathbf{e}$. *The expected value of* $g(\mathbf{e})$ *is* $E[g(\mathbf{e})] = \sum_{\mathbf{n}} c_{\mathbf{n}} \Phi_X(\mathbf{Un})$ *where* $c_{\mathbf{n}} = \frac{1}{|det\mathbf{V}|} \int_{\mathcal{P}_0} g(\mathbf{e}) e^{-j\mathbf{e}^T\mathbf{Un}} d\mathbf{e}$. ♢

Proof.

$$\begin{aligned} E[g(\mathbf{e})] &= \int_{\mathcal{P}_0} g(\mathbf{e}) f_{\mathbf{E}}(\mathbf{e}) d\mathbf{e} \\ &= \int_{\mathcal{P}_0} g(\mathbf{e}) \sum_{\mathbf{n}} f_{\mathbf{X}}(\mathbf{e} + \mathbf{Vn}) d\mathbf{e} \quad \text{(from (4.10))} \\ &= \int_{\mathcal{P}_0} g(\mathbf{e}) \frac{1}{|det\mathbf{V}|} \sum_{\mathbf{n}} \Phi_{\mathbf{X}}(\mathbf{Un}) e^{-j\mathbf{e}^T\mathbf{Un}} d\mathbf{e} \quad \text{(from (4.11))} \\ &= \sum_{\mathbf{n}} \Phi_{\mathbf{X}}(\mathbf{Un}) \frac{1}{|det\mathbf{V}|} \int_{\mathcal{P}_0} g(\mathbf{e}) e^{-j\mathbf{e}^T\mathbf{Un}} d\mathbf{e} \\ &= \sum_{\mathbf{n}} c_{\mathbf{n}} \Phi_X(\mathbf{Un}) \end{aligned} \tag{4.16}$$

as claimed. Note that $g(\mathbf{e})$ and therefore $c_{\mathbf{n}}$ can be a vector or even a matrix. We assume that interchanging the infinite sum and the integral in the above proof is permissible. ■

Corollary 1: Error moments. For any random vector $\mathbf{x}$, the first and second order

moments of the quantization error $\mathbf{e}$ of a lattice quantizer $Q(VOR, \mathbf{V})$ are:

$$E[\mathbf{e}] = \sum_{\mathbf{n}\neq\mathbf{0}} \mathbf{c_n}\Phi_X(\mathbf{Un}), \qquad E[\mathbf{ee}^T] = \mathbf{G}_D(VOR, \mathbf{V}) + \sum_{\mathbf{n}\neq\mathbf{0}} \mathbf{C_n}\Phi_{\mathbf{X}}(\mathbf{Un}) \tag{4.17}$$

where

$$\mathbf{c_n} = \frac{1}{|det\mathbf{V}|}\int_{VOR(\mathbf{V})} \mathbf{e}\; e^{-j\mathbf{e}^T\mathbf{Un}}d\mathbf{e}, \qquad \mathbf{C_n} = \frac{1}{|det\mathbf{V}|}\int_{VOR(\mathbf{V})} \mathbf{ee}^T e^{-j\mathbf{e}^T\mathbf{Un}}d\mathbf{e}. \tag{4.18}$$

$\diamondsuit$

Proof. Apply Theorem 3 with $\mathcal{P}_0 = VOR(\mathbf{V})$ to $g(\mathbf{e}) = \mathbf{e}$ and $g(\mathbf{e}) = \mathbf{ee}^T$ respectively. Because of the symmetry of $VOR(\mathbf{V})$, $\mathbf{c_0} = \mathbf{0}$. Moreover, $\mathbf{C_0}$ is what we defined as $\mathbf{G}_D(VOR, \mathbf{V})$ in (4.6). ■

Note that if $\mathbf{x}$ is Nyquist-$\mathbf{V}$, all the terms of the infinite summations in (4.17) vanish in view of Theorem 1.

4.3 Subtractive Dithering

Let $\mathbf{v}$ be a random vector, statistically independent of the input vector $\mathbf{x}$. Adding this so called dither vector before the quantization and subtracting after it, we have the subtractive-dithered lattice quantization, as depicted in Fig. 4.6.

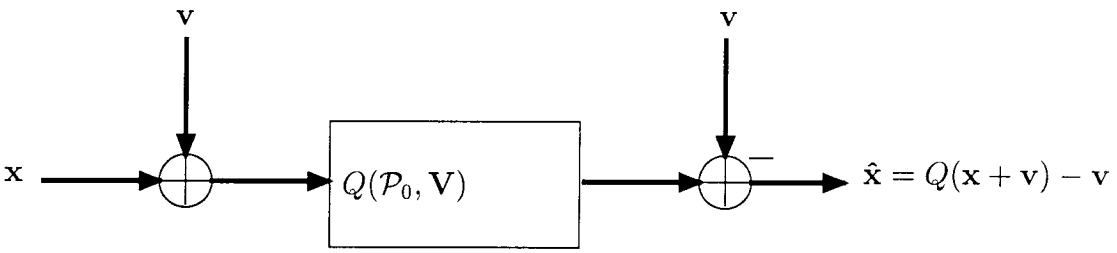


Fig. 4.6: Subtractively dithered lattice quantizer.

The error vector is $\mathbf{e} = \mathbf{x} - (Q(\mathbf{x}+\mathbf{v}) - \mathbf{v}) = \mathbf{x} + \mathbf{v} - Q(\mathbf{x}+\mathbf{v})$. Notice that this error is the same as the conventional quantization error for an input vector $\mathbf{x}+\mathbf{v}$. The characteristic function of the sum $\mathbf{x}+\mathbf{v}$ is $\Phi_{\mathbf{X}}(\Omega)\Phi_{\mathbf{V}}(\Omega)$ where $\Phi_{\mathbf{X}}(\Omega)$ and $\Phi_{\mathbf{V}}(\Omega)$ are the characteristic functions of $\mathbf{x}$ and $\mathbf{v}$ respectively. Hence $\mathbf{x}+\mathbf{v}$ is Nyquist-$\mathbf{V}$

whenever $\mathbf{v}$ is Nyquist-$\mathbf{V}$ and from Theorem 1, it follows that the quantization error is uniform in the quantization basic cell, $\mathcal{P}_0$. However, more is true as the following theorem shows:

Theorem 4. *In the subtractive quantization scheme of Fig. 4.6, the error vector* $\mathbf{e}$ *is statistically independent of the input vector* $\mathbf{x}$ *and uniformly distributed in* $\mathcal{P}_0$ *if and only if the dither* $\mathbf{v}$ *is Nyquist-*$\mathbf{V}$*, that is* $\Phi_{\mathbf{V}}(\mathbf{Un}) = \delta(\mathbf{n})$*, where* $\mathbf{U} = 2\pi\mathbf{V}^{-T}$. $\diamondsuit$

Proof. Let $\mathbf{u} = \mathbf{x}+\mathbf{v}$. The conditional density of $\mathbf{u}$, conditioned on $\mathbf{x}$, is $f_{\mathbf{U/X}}(\mathbf{u/x}) = f_{\mathbf{V}}(\mathbf{u}-\mathbf{x})$ and the corresponding characteristic function is $\Phi_{\mathbf{U/X}}(\mathbf{\Omega}) = \Phi_{\mathbf{V}}(\mathbf{\Omega})e^{j\mathbf{\Omega}^T\mathbf{x}}$. Hence using (4.11), we can write the conditional density function of the error vector as:

$$f_{\mathbf{E/X}}(\mathbf{e/x}) = \frac{1}{|det\mathbf{V}|}\sum_{\mathbf{n}}\Phi_{\mathbf{U/X}}(\mathbf{Un})e^{-j\mathbf{e}^T\mathbf{Un}} = \frac{1}{|det\mathbf{V}|}\sum_{\mathbf{n}}\Phi_{\mathbf{V}}(\mathbf{Un})e^{j\mathbf{x}^T\mathbf{Un}}e^{-j\mathbf{e}^T\mathbf{Un}} \tag{4.19}$$

for $\mathbf{e} \in \mathcal{P}_0$ and 0 elsewhere. One can think of this as the nonseparable discrete Fourier transform of the sequence $\Phi_{\mathbf{V}}(\mathbf{Un})e^{-j\mathbf{e}^T\mathbf{Un}}$, $\mathbf{x}$ being the transform domain vector. Hence from the uniqueness of Fourier transform, this is independent of $\mathbf{x}$ if and only if $\Phi_{\mathbf{V}}(\mathbf{Un})e^{-j\mathbf{e}^T\mathbf{Un}} = \delta(\mathbf{n})$ which is equivalent to $\Phi_{\mathbf{V}}(\mathbf{Un}) = \delta(\mathbf{n})$. ■

If $\mathcal{P}_0 = VOR(\mathbf{V})$ and the condition of the theorem is satisfied, then $E[\mathbf{e}] = \mathbf{0}$ and $E[\mathbf{ee}^T] = \mathbf{G}_D(VOR, \mathbf{V})$, where $\mathbf{G}_D(VOR, \mathbf{V})$ is defined as in (4.6).

4.3.1 Nyquist-V Dither Vectors: Examples and Generation

In Theorem 2, we provided some classes of random vectors that are Nyquist-$\mathbf{V}$. Any such vector will serve as a dither vector as long as it is independent of the input vector $\mathbf{x}$. In particular, as given in Example 1, we can use a dither vector that is uniform in $SPD(\mathbf{V})$ or $VOR(\mathbf{V})$. The one that is uniform in $SPD(\mathbf{V})$ is relatively simple to generate and a method for generating such a dither is given next.

Generation of a Nyquist-V vector: We will show how to obtain a random vector that is uniform in $SPD(\mathbf{V})$ and therefore is Nyquist-$\mathbf{V}$. First generate a set of D

independent random variables $z_1, z_2, .., z_D$ each of which is uniform in [-1/2, 1/2). Form the vector $\mathbf{z} = [\, z_1 \quad z_2 \quad \ldots \quad z_D \,]^T$. The vector $\mathbf{v} = \mathbf{V}\mathbf{z}$ is Nyquist-$\mathbf{V}$ because:

$$\Phi_{\mathbf{V}}(\mathbf{Un}) = E_{\mathbf{V}}[e^{j\mathbf{v}^T\mathbf{Un}}] = E_{\mathbf{Z}}[e^{j\mathbf{z}^T\mathbf{V}^T\mathbf{Un}}] = E_{\mathbf{Z}}[e^{j2\pi\mathbf{z}^T\mathbf{n}}] = \delta(\mathbf{n}). \tag{4.20}$$

The procedure is illustrated in Fig. 4.7 for $D = 2$. Since the error vector of a lattice

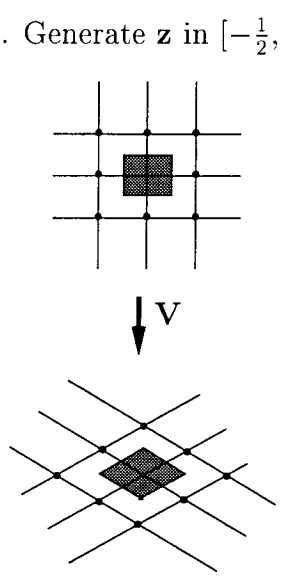


Fig. 4.7: Generation of a Nyquist-$\mathbf{V}$ vector in $2 - D$.

quantizer $Q(\mathcal{P}_0, \mathbf{V})$ can be made uniform in $\mathcal{P}_0$ by applying a Nyquist dither, we will give our attention to the moments of that error. All of the results stated below can actually be viewed as the properties of the underlying lattice, but the reader should keep in mind that they will become the properties of the quantization error if the input is Nyquist, or if a Nyquist-dither is added to the input prior to quantization.

4.3.2 Performance Comparison of Lattice Quantizers

Note that $\mathbf{G}_D(\mathcal{P}_0, \mathbf{V})$, the second moment of the error of a lattice quantizer $Q(\mathcal{P}_0, \mathbf{V})$ with Nyquist-$\mathbf{V}$ input, is a positive definite symmetric matrix. The total mean square error of the quantizer is the trace of this matrix: $E[\|\mathbf{e}\|^2] = Tr(\mathbf{G}_D(\mathcal{P}_0, \mathbf{V}))$.

Orthogonal lattices: An orthogonal lattice is a lattice whose generator matrix $\mathbf{V}$ satisfies

$$\mathbf{V}\mathbf{V}^T = |det\mathbf{V}|^{2/D}\mathbf{\Lambda}, \tag{4.21}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with diagonal elements $\lambda_i > 0$. To preserve the determinants of both sides of (4.21), we have $\prod_{i=1}^{D} \lambda_i = 1$.

Notice that an orthogonal lattice quantizer with $VOR(\mathbf{V})$ as its basic cell can be considered as a collection of scalar uniform quantizers for each dimension with possibly different step sizes. We note the following result on the second moment matrix, $\mathbf{G}_D(VOR, \mathbf{V})$, of an orthogonal lattice quantizer $Q(VOR, \mathbf{V})$:

Fact 1. If the lattice $\mathcal{L}(\mathbf{V})$ is orthogonal, that is $\mathbf{V}\mathbf{V}^T = |det\mathbf{V}|^{2/D}\mathbf{\Lambda}$, then

$$\mathbf{G}_D(VOR, \mathbf{V}) = \frac{1}{12}|det\mathbf{V}|^{2/D}\mathbf{\Lambda}. \tag{4.22}$$

$\diamondsuit$

See Appendix B for the proof. As a special case, if $\mathbf{V}\mathbf{V}^T = |det\mathbf{V}|^{2/D}\mathbf{I}$, then $\mathbf{G}_D(VOR, \mathbf{V}) = \frac{1}{12}|det\mathbf{V}|^{2/D}\mathbf{I}$, and therefore $\frac{1}{D}E[\|\mathbf{e}\|^2] = \frac{1}{12}|det\mathbf{V}|^{2/D}$. Taking this as a reference, we can compare the performances of other lattice quantizers. We will normalize the total mean square error per dimension of any lattice quantizer $Q(\mathcal{P}_0, \mathbf{V})$ by $|det\mathbf{V}|^{2/D}$, giving a proper figure of merit for lattices of different volume and dimension D.

Definition 8. The *dimensionless second moment* of a lattice quantizer $Q(\mathcal{P}_0, \mathbf{V})$, denoted by $\sigma_D^2(\mathcal{P}_0, \mathbf{V})$, is defined as

$$\sigma_D^2(\mathcal{P}_0, \mathbf{V}) = \frac{1}{D|detV|^{2/D}} Tr(\mathbf{G}_D(\mathcal{P}_0, \mathbf{V})) = \frac{1}{D|det\mathbf{V}|^{1+2/D}} \int_{\mathcal{P}_0} \|\mathbf{e}\|^2 d\mathbf{e} \tag{4.23}$$

where $\mathbf{G}_D(\mathcal{P}_0, \mathbf{V})$ is as in (4.6). $\diamondsuit$

The quantity $\sigma_D^2(\mathcal{P}_0, \mathbf{V})$ also comes out of high bit rate analysis of lattice quantizers [LZ94, Ger79, LZ94]. It is proven in [LZ94] that for an undithered lattice quantizer, as the unit volume, $|detV|$ of a quantizer $Q(\mathcal{P}_0, \mathbf{V})$ goes to 0, the normalized mean square error approaches the limit $\sigma_D^2(\mathcal{P}_0, \mathbf{V})$. The name *dimensionless second moment* is used in [LZ94].

The following fact is on the performance of orthogonal lattice quantizers. At best, they can perform as good as a collection of independent scalar quantizers with identical step sizes. The reader is referred to Appendix B for the proof.

Fact 2. Let $Q(\mathcal{P}_0, \mathbf{V})$ be an orthogonal lattice quantizer; that is let the generator matrix $\mathbf{V}$ satisfy (4.21). Then,

$$\sigma_D^2(\mathcal{P}_0, \mathbf{V}) \geq \frac{1}{12}, \tag{4.24}$$

with equality if and only if $\mathbf{V}\mathbf{V}^T = |det\mathbf{V}|^{2/D}\mathbf{I}$ and $\mathcal{P}_0 = VOR(\mathbf{V})$. $\diamondsuit$

The following result is on the performance of lattice quantizers whose quantization basic cells are $SPD(\mathbf{V})$ rather than $VOR(\mathbf{V})$. Note that this result is not a special case of the previous one, where we assumed $\mathbf{V}$ was orthogonal. Here, there is no assumption on $\mathbf{V}$.

Fact 3. Given a lattice generator matrix $\mathbf{V}$,

$$\sigma_D^2(SPD, \mathbf{V}) \geq \frac{1}{12}, \tag{4.25}$$

with equality if and only if $\mathbf{V}\mathbf{V}^T = |det\mathbf{V}|^{2/D}\mathbf{I}$. $\diamondsuit$

Proof. By making a change of variable as we did in the proof of Fact 1, it is easy to see that $\mathbf{G}_D(SPD, \mathbf{V}) = \frac{1}{12}\mathbf{V}\mathbf{V}^T$. Hence,

$$\begin{aligned}
\sigma_D^2(SPD) &= \frac{1}{D|detV|^{2/D}} Tr(\frac{1}{12}\mathbf{V}\mathbf{V}^T) \\
&\geq \frac{1}{12|detV|^{2/D}} |det\mathbf{V}\mathbf{V}^T|^{1/D} \quad \text{(see below)} \\
&= \frac{1}{12}
\end{aligned} \tag{4.26}$$

The inequality follows from the AM-GM inequality and the Hadamard inequality [Vai93] as explained next. The diagonal elements of the positive definite matrix $\mathbf{V}\mathbf{V}^T$ are positive. Hence, their arithmetic mean is greater than or equal to their geometric mean. And by the Hadamard inequality, the product of the diagonal elements is

greater than or equal to the determinant of $\mathbf{V}\mathbf{V}^T$. The former is an equality if and only if the diagonal elements of $\mathbf{V}\mathbf{V}^T$ are the same and the latter is an equality if and only if $\mathbf{V}\mathbf{V}^T$ is diagonal. Hence, the result follows. ■

As we noted before, for a given lattice $\mathcal{L}(\mathbf{V})$, the minimum dimensionless second moment is achieved by the basic cell $VOR(\mathbf{V})$. One can ask the question: among all the lattices in R^D, what is the optimum lattice that will minimize the dimensionless second moment $\sigma_D^2(VOR, \mathbf{V})$? This question turns out to be theoretically very challenging. The answer is not known for arbitrary D and there is no proof of optimality for dimensions higher than 3 (see for example, [GS88]).

Examples of optimum lattices. Here are some lattices that have minimum dimensionless second moments:

Case where $D = 1$. The only lattice is the points of the form $\Delta n, \forall n \in Z$, $\Delta \in R$. Any basic cell $\mathcal{P}$ has a total length of Δ. Obviously, the minimum dimensionless second moment is achieved by $VOR(\Delta) = [-\frac{\Delta}{2}, \frac{\Delta}{2})$ and its value is

$$\sigma_1^2(VOR, \Delta) = \frac{1}{12} \tag{4.27}$$

Case where $D = 2$. The optimum lattice that minimizes $\sigma_D^2(VOR, \mathbf{V})$ is the one whose $VOR(\mathbf{V})$ is the regular hexagon [Ger79]. A generating matrix for this lattice is

$$\mathbf{V} = \begin{pmatrix} 3 & \frac{3}{2} \\ 0 & \frac{\sqrt{3}}{2} \end{pmatrix} \tag{4.28}$$

The unit volume of the lattice is: $|det\mathbf{V}| = \frac{3\sqrt{3}}{2}$. By explicitly evaluating integrals, we have:

$$E[\mathbf{e}\mathbf{e}^T] = \mathbf{G}_2(VOR, \mathbf{V}) = \frac{5}{24}\mathbf{I}, \tag{4.29}$$

where $\mathbf{I}$ is 2×2 identity matrix. The corresponding dimensionless second moment is:

$$\sigma_2^2(VOR, \mathbf{V}) = \frac{5}{24} / \frac{3\sqrt{3}}{2} = \frac{5}{36\sqrt{3}} \simeq 0.08018754 \tag{4.30}$$

Compare this to that of optimum one-dimensional lattice:

$$\sigma_1^2(VOR, \Delta) = \frac{1}{12} = 0.0833.. \tag{4.31}$$

Case where $D = 3$. The optimum lattice is the body-centered cubic lattice, also called the truncated octahedron as is proven by Barnes and Sloane [BS83]. This lattice has

$$\sigma_3^2(VOR, \mathbf{V}) = \frac{19}{192\sqrt[3]{2}} \simeq 0.0785433.. \tag{4.32}$$

Case where $D = \infty$. The limiting value of minimum $\sigma_D^2(VOR, \mathbf{V})$ is [LZ94],

$$\liminf_{D\to\infty} \sigma_D^2 = \frac{1}{2\pi e} \simeq 0.058823... \tag{4.33}$$

For a tabulation of lattices that have best known $\sigma_D^2(VOR, \mathbf{V})$, see [LZ94].

After the observation in (4.29) that $\mathbf{G}_2(VOR, \mathbf{V})$ is diagonal with equal elements, these authors suspected that this might be true for any optimum lattice of arbitrary dimension. This turns out to be indeed the case, as elaborated in the next theorem. Assume the dimension D is given and we look at different lattices with the objective of minimizing the dimensionless second moment $\sigma_D^2(\mathcal{P}_0, \mathbf{V})$. Hence the quantization basic cells are chosen to be $VOR(\mathbf{V})$ for each lattice generator matrix $\mathbf{V}$. We have the following result:

Theorem 5. *For a lattice quantizer $Q(\mathcal{P}_0, \mathbf{V})$ to be optimum, that is, to have the minimum dimensionless second moment $\sigma_D^2(\mathcal{P}_0, \mathbf{V})$, it is necessary that $\mathcal{P}_0 = VOR(\mathbf{V})$, and*

$$\mathbf{G}_D(VOR, \mathbf{V}) = c\mathbf{I} \tag{4.34}$$

where $c = \sigma_D^2(VOR, \mathbf{V})|det\mathbf{V}|^{2/D}$ and $\mathbf{I}$ is the $D \times D$ identity matrix. $\diamondsuit$

Comment. Note that $\mathbf{G}_D(\mathcal{P}_0, \mathbf{V})$ is the second moment matrix of a vector $\mathbf{e}$ with uniform pdf in $\mathcal{P}_0$. By Theorem 1, uniformity of the error in $\mathcal{P}_0$ is equivalent to

the Nyquist-$\mathbf{V}$ condition on the input vector $\mathbf{x}$. This can be assured by adding an independent Nyquist-$\mathbf{V}$ dither, as seen from Theorem 4.

During the preparation of this chapter, the authors noticed that this result has appeared very recently in [ZF95] and a very similar proof has been provided in [ZF94]. Nevertheless, we provide our proof here for completeness and convenience.

Proof. As we noted before, for any given $\mathbf{V}$, the minimum dimensionless second moment is achieved by the quantization basic cell $VOR(\mathbf{V})$. Hence we take $\mathcal{P}_0 = VOR(\mathbf{V})$. Define a new random vector $\mathbf{z} = \mathbf{Q}^{-1}\mathbf{x}$ for some nonsingular $\mathbf{Q}$, and consider Fig. 4.8. Since $\hat{\mathbf{x}}$ is on the lattice $\mathcal{L}(\mathbf{V})$, the vector $\hat{\mathbf{z}}$ is on the lattice $\mathcal{L}(\mathbf{Q}^{-1}\mathbf{V})$. We

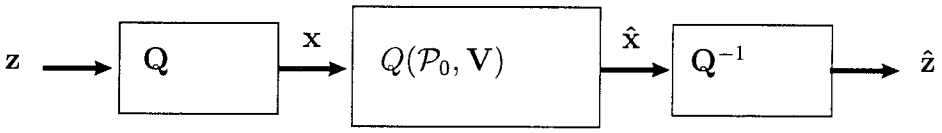


Fig. 4.8: Transformation of a lattice quantizer. Here, $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$ is uniform in $VOR(\mathbf{V})$. This can be assured by subtractive dithering. Hence, $\mathbf{f} = \mathbf{z} - \hat{\mathbf{z}}$ is uniform in a basic cell of the transformed lattice.

can therefore regard Fig. 4.8 as a lattice quantizer for the vector $\mathbf{z}$, with the quantized values on $\mathcal{L}(\mathbf{Q}^{-1}\mathbf{V})$. Define the quantization errors $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$ and $\mathbf{f} = \mathbf{z} - \hat{\mathbf{z}}$. Then $\mathbf{f} = \mathbf{Q}^{-1}\mathbf{e}$. Since $\mathbf{e}$ is uniform in $VOR(\mathbf{V})$, the error $\mathbf{f}$ is uniform in a basic cell, $\mathcal{P}$ of $\mathcal{L}(\mathbf{Q}^{-1}\mathbf{V})$. Assuming that $\mathbf{V}$ is optimal for the dimension D , the dimensionless second moments should satisfy

$$\sigma_D^2(\mathcal{P}, \mathbf{Q}^{-1}\mathbf{V}) \geq \sigma_D^2(VOR, \mathbf{V}). \tag{4.35}$$

Observe that $E[\mathbf{f}\mathbf{f}^T] = \mathbf{Q}^{-1}E[\mathbf{e}\mathbf{e}^T]\mathbf{Q}^{-T}$. Let us choose $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{Q}^T = E[\mathbf{e}\mathbf{e}^T]$, so $E[\mathbf{f}\mathbf{f}^T] = \mathbf{I}$. Substituting the expressions

$$\sigma_D^2(\mathcal{P}, \mathbf{Q}^{-1}\mathbf{V}) = \frac{E[\|\mathbf{f}\|^2]}{D|det\mathbf{Q}^{-1}\mathbf{V}|^{2/D}}, \quad \text{and} \quad \sigma_D^2(VOR, \mathbf{V}) = \frac{E[\|\mathbf{e}\|^2]}{D|det\mathbf{V}|^{2/D}} \tag{4.36}$$

into (4.35), we can simplify it to

$$|det\mathbf{Q}\mathbf{Q}^T|^{1/D} \geq \frac{1}{D}Tr(\mathbf{Q}\mathbf{Q}^T). \tag{4.37}$$

Let λ_i be the eigenvalues of the Hermitian matrix $\mathbf{QQ}^T$. Hence the determinant and the trace above are, respectively, the product and the sum of these eigenvalues. So the preceding equation is equivalent to $(\prod_{i=1}^{D} \lambda_i)^{1/D} \geq \frac{1}{D} \sum_{i=1}^{D} \lambda_i$. Since by construction $\mathbf{QQ}^T$ is positive definite, $\lambda_i > 0$ for all i. We can therefore apply the AM-GM inequality to conclude $\frac{1}{D} \sum_{i=1}^{D} \lambda_i \geq (\prod_{i=1}^{D} \lambda_i)^{1/D}$. The preceding two inequalities on $\{\lambda_i\}$ can be simultaneously true if and only if λ_i is identical for all i. Since $\mathbf{QQ}^T$ is Hermitian, this proves that $\mathbf{QQ}^T = \lambda \mathbf{I}$. So we have proved that $E[\mathbf{ee}^T] = \lambda \mathbf{I}$. Combining this with the definition of $\sigma_D^2(VOR, \mathbf{V})$, we obtain (4.34) indeed. ■

4.4 Nonsubtractive Dithering

In subtractive dithering, one should regenerate the dither vector exactly at the reconstruction end. This is, in most cases, undesirable. The easiest remedy is not to subtract the dither vector, and this results in the nonsubtractive dithering scheme. Referring to Fig. 4.9, we define the error vector to be $\mathbf{e} = \mathbf{x} - Q(\mathbf{x} + \mathbf{v})$. The error is

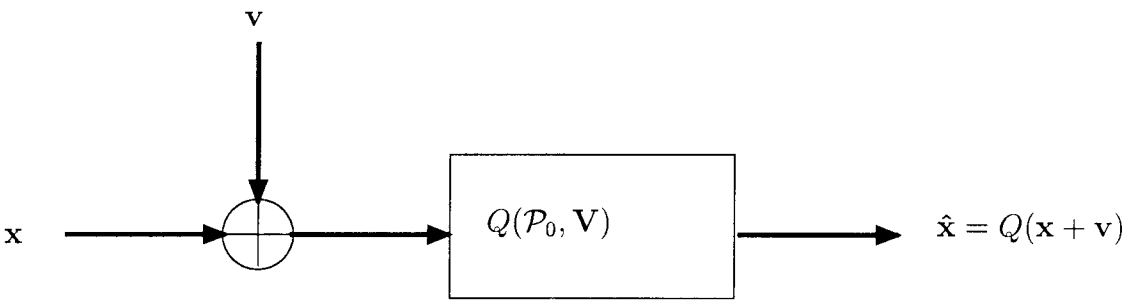


Fig. 4.9: Nonsubtractively dithered lattice quantizer.

no longer a periodic function of the input and therefore we do not have a periodical relationship between the error and the input pdf's similar to (4.10) or (4.11). Hence, as can be shown, the error cannot be rendered statistically independent from the input. However, the moments of the error can be rendered independent from the input as will be elaborated next. This result is the generalization of the well-known one-dimensional nonsubtractive dithering result [LWV92], [GS93]. First we will give a lemma that will

express the relevant moments in terms of gradients of a function of dither.

Let ∇ and $\nabla\nabla^T$ denote the first and second order gradient operators operating on functions of D variables, $\omega_1, \omega_2, .., \omega_D$:

$$\nabla = [\ \frac{\partial}{\partial\omega_1} \quad \frac{\partial}{\partial\omega_2} \cdots \quad \frac{\partial}{\partial\omega_D}\ ]^T, \qquad (\nabla\nabla^T)_{ij} = \frac{\partial^2}{\partial\omega_i\partial\omega_j} \tag{4.38}$$

Let $\mathbf{z}$ be a random vector that is uniform in the quantization basic cell $\mathcal{P}_0$ of the lattice quantizer $Q(\mathcal{P}_0, \mathbf{V})$. Let $f_{\mathbf{V}}(\mathbf{v})$ and $f_{\mathbf{Z}}(\mathbf{z})$ be the pdf's of $\mathbf{v}$ and $\mathbf{z}$ respectively. By definition,

$$f_{\mathbf{Z}}(\mathbf{z}) = \begin{cases} \frac{1}{|det\mathbf{V}|}, & \mathbf{z} \in \mathcal{P}_0; \\ \\ 0 & \text{elsewhere} \end{cases} \tag{4.39}$$

Lemma 1. The first and second moments of the error vector $\mathbf{e}$ of a nonsubtractively dithered lattice quantizer $Q(\mathcal{P}_0, \mathbf{V})$ conditioned on the input vector $\mathbf{x}$ are:

$$E[\mathbf{e}/\mathbf{x}] = \frac{1}{j}\sum_{\mathbf{n}} \nabla H(\mathbf{Un}) e^{j\mathbf{x}^T\mathbf{Un}} \tag{4.40}$$

$$E[\mathbf{e}\mathbf{e}^T/\mathbf{x}] = \frac{1}{j^2}\sum_{\mathbf{n}} \nabla\nabla^T H(\mathbf{Un}) e^{j\mathbf{x}^T\mathbf{Un}} \tag{4.41}$$

where,

$$H(\mathbf{\Omega}) = \int h(\mathbf{x}) e^{-j\mathbf{\Omega}^T\mathbf{x}} d\mathbf{x} \qquad h(\mathbf{e}) = f_{\mathbf{V}}(\mathbf{e}) * f_{\mathbf{Z}}(-\mathbf{e}) \tag{4.42}$$

$\diamondsuit$

Remark. Note that the extension of the above result to higher moments is straightforward by defining the corresponding operators in an obvious way. However, our interest will only be in the first and second order moments.

Proof. Since we do not subtract the dither after the quantizer, the reproduction points are the lattice points of the form $\mathbf{Vn}$. That is, $Q(\mathbf{x}+\mathbf{v}) = \mathbf{Vn}$ for some $\mathbf{n} \in Z^D$. Hence, the corresponding error vector is $\mathbf{e} = \mathbf{x} - \mathbf{Vn}$. Note that, given $\mathbf{x}$, this is a discrete random vector. It has the probability mass function:

$$P_{\mathbf{E}/\mathbf{X}}(\mathbf{x} - \mathbf{Vn}) \ = \ Prob\{Q(\mathbf{x}+\mathbf{v}) = \mathbf{Vn}\}$$

$$\begin{aligned} &= Prob\{\mathbf{x}+\mathbf{v}=\mathbf{x}_0+\mathbf{Vn}, \mathbf{x}_0 \in \mathcal{P}_0\} \\ &= \int_{\mathcal{P}_0(\mathbf{Vn}-\mathbf{x})} f_{\mathbf{V}}(\mathbf{v})d\mathbf{v} \end{aligned} \tag{4.43}$$

where $\mathcal{P}_0(\mathbf{Vn}-\mathbf{x})$ denotes the translated region of $\mathcal{P}_0$ by the vector $\mathbf{Vn}-\mathbf{x}$. Using the artificial random vector $\mathbf{z}$ defined by the pdf in (4.39), one can express the preceding as a convolution:

$$P_{\mathbf{E/X}}(\mathbf{x}-\mathbf{Vn}) = |det\mathbf{V}| \int f_{\mathbf{V}}(\mathbf{v}) f_{\mathbf{Z}}(\mathbf{v}-\mathbf{Vn}+\mathbf{x})d\mathbf{v} \tag{4.44}$$

Hence,

$$\begin{aligned} P_{\mathbf{E/X}}(\mathbf{e}) &= |det\mathbf{V}| \int f_{\mathbf{V}}(\mathbf{v}) f_{\mathbf{Z}}(\mathbf{v}+\mathbf{e})d\mathbf{v} \\ &= |det\mathbf{V}|\ h(-\mathbf{e}) \end{aligned} \tag{4.45}$$

where

$$h(\mathbf{e}) = f_{\mathbf{V}}(\mathbf{e}) * f_{\mathbf{Z}}(-\mathbf{e}) \tag{4.46}$$

Now, the first order moment of the error vector is:

$$\begin{aligned} E[\mathbf{e/x}] &= \sum \mathbf{e} P_{\mathbf{E/X}}(\mathbf{e}) \\ &= \sum_{\mathbf{n}} (\mathbf{x}-\mathbf{Vn}) |det\mathbf{V}| h(-\mathbf{x}+\mathbf{Vn}) \\ &= \sum_{\mathbf{n}} g(\mathbf{x}+\mathbf{Vn}) \end{aligned} \tag{4.47}$$

where $g(\mathbf{x})$ is defined as $|det\mathbf{V}|\mathbf{x}h(-\mathbf{x})$. The Fourier transform of $g(\mathbf{x})$ is $G(\mathbf{\Omega}) = \frac{1}{j}|det\mathbf{V}|\nabla H(-\mathbf{\Omega})$, where $H(\mathbf{\Omega})$ is the Fourier transform of $h(\mathbf{e})$, that is

$$H(\mathbf{\Omega}) = \Phi_{\mathbf{V}}(-\mathbf{\Omega})\Phi_{\mathbf{Z}}(\mathbf{\Omega}). \tag{4.48}$$

By using the Fourier series representation (see Appendix A), one can write (4.47) as:

$$E[\mathbf{e}/\mathbf{x}] = \frac{1}{|det\mathbf{V}|} \sum_{\mathbf{n}} G(\mathbf{Un}) e^{-j\mathbf{x}^T\mathbf{Un}} \tag{4.49}$$

which reduces to (4.40). The derivation of (4.41) is through the same steps and is omitted. ■

Using these results and noting the uniqueness property of Fourier series, the next theorem follows:

Theorem 6. *Consider the nonsubtractive quantization scheme of Fig. 4.9. Let* $H(\mathbf{\Omega})$ *be as in (4.48).*

1. The first order moment of the error vector is independent of the input if and only if $\nabla H(\mathbf{\Omega})$ *is Nyquist-*$\mathbf{U}$*, that is* $\nabla H(\mathbf{Un}) = \mathbf{c}\ \delta(\mathbf{n})$,

2. The second order moment matrix of the error vector is independent of the input if and only if $\nabla\nabla^T H(\mathbf{\Omega})$ *is Nyquist-*$\mathbf{U}$*, that is* $\nabla\nabla^T H(\mathbf{Un}) = \mathbf{C}\ \delta(\mathbf{n})$.

If the corresponding conditions are satisfied, then

$$E[\mathbf{e}/\mathbf{x}] = E[\mathbf{e}] = E[\mathbf{z}] - E[\mathbf{v}], \quad E[\mathbf{e}\mathbf{e}^T/\mathbf{x}] = E[\mathbf{e}\mathbf{e}^T] = E[(\mathbf{z}-\mathbf{v})(\mathbf{z}-\mathbf{v})^T] \tag{4.50}$$

respectively, where $\mathbf{z}$ *is uniform in* $\mathcal{P}_0$ *and independent of* $\mathbf{v}$. ♢

Remark. If the conditions are satisfied with a symmetric basic cell $\mathcal{P}_0$, then $E[\mathbf{e}] = -E[\mathbf{v}]$, and $E[\mathbf{e}\mathbf{e}^T] = \mathbf{G}_D(\mathcal{P}_0, \mathbf{V}) + E[\mathbf{v}\mathbf{v}^T]$, where $\mathbf{G}_D(\mathcal{P}_0, \mathbf{V})$ is defined as in (4.6). In particular, if $\mathcal{P}_0 = VOR(\mathbf{V})$, then $E[\mathbf{e}\mathbf{e}^T] = \mathbf{G}_D(VOR, \mathbf{V}) + E[\mathbf{v}\mathbf{v}^T]$.

Proof. The necessary and sufficient conditions follow from Lemma 1. If the corresponding conditions are satisfied, then

$$E[\mathbf{e}] = \frac{1}{j}\nabla H(\mathbf{0}), \quad \text{and} \quad E[\mathbf{e}\mathbf{e}^T] = \frac{1}{j^2}\nabla\nabla^T H(\mathbf{0}), \tag{4.51}$$

respectively. Now, by (4.46), $\mathrm{h}(\mathbf{e})$ can be considered as the pdf of a random vector $\mathbf{v} - \mathbf{z}$, where $\mathbf{z}$ is independent from $\mathbf{v}$ and uniform in $\mathcal{P}_0$. Hence, from the moment

generating property of characteristic functions, we have

$$\frac{1}{j}\nabla H(\mathbf{0}) = E[\mathbf{z}-\mathbf{v}], \quad \text{and} \quad \frac{1}{j^2}\nabla\nabla^T H(\mathbf{0}) = E[(\mathbf{z}-\mathbf{v})(\mathbf{z}-\mathbf{v})^T]. \tag{4.52}$$

■

Example 2. Let $\mathbf{v}$ be any Nyquist-$\mathbf{V}$ random vector, that is, $\Phi_{\mathbf{V}}(\mathbf{Un}) = \delta(\mathbf{n})$. Then the condition for the first part of the theorem is satisfied. To see this:

$$\nabla H(\mathbf{\Omega}) = \Phi_{\mathbf{V}}(-\mathbf{\Omega})\nabla\Phi_{\mathbf{Z}}(\mathbf{\Omega}) - \Phi_{\mathbf{Z}}(\mathbf{\Omega})\nabla\Phi_{\mathbf{V}}(-\mathbf{\Omega}) \tag{4.53}$$

Since $\Phi_{\mathbf{Z}}$ itself is Nyquist and $\Phi_{\mathbf{V}}$ is chosen to be so, ∇H is Nyquist as well. Hence $E[\mathbf{e}/\mathbf{x}] = E[\mathbf{e}] = E[\mathbf{z}] - E\mathbf{v}]$. This is zero if (i) the dither is uniform in the quantization basic cell or (ii) the dither is uniform in any symmetric basic cell and the quantization basic cell is symmetric. The dither vector that is uniform in $SPD(\mathbf{V})$ satisfies the condition of the theorem and it produces zero-mean error if the quantization basic cell is symmetric.

Example 3. Let $\mathbf{v} = \mathbf{z}_1 + \mathbf{z}_2$ where $\mathbf{z}_1$ and $\mathbf{z}_2$ are independent random vectors each of which is Nyquist-$\mathbf{V}$. Then the condition for the second part of the theorem is satisfied, because:

$$\begin{aligned}
\nabla\nabla^T H(\mathbf{\Omega}) &= \nabla(\Phi_{\mathbf{V}}(-\mathbf{\Omega})\nabla^T\Phi_{\mathbf{Z}}(\mathbf{\Omega}) - \Phi_{\mathbf{Z}}(\mathbf{\Omega})\nabla^T\Phi_{\mathbf{V}}(-\mathbf{\Omega})) \\
&= \Phi_{\mathbf{V}}(-\mathbf{\Omega})\nabla\nabla^T\Phi_{\mathbf{Z}}(\mathbf{\Omega}) - \nabla\Phi_{\mathbf{V}}(-\mathbf{\Omega})\nabla^T\Phi_{\mathbf{Z}}(\mathbf{\Omega}) \\
&\quad -\nabla\Phi_{\mathbf{Z}}(\mathbf{\Omega})\nabla^T\Phi_{\mathbf{V}}(-\mathbf{\Omega}) + \Phi_{\mathbf{Z}}(\mathbf{\Omega})\nabla\nabla^T\Phi_{\mathbf{V}}(-\mathbf{\Omega})
\end{aligned} \tag{4.54}$$

The first term is Nyquist because $\Phi_{\mathbf{V}}$, being the product of two Nyquist functions, is Nyquist. From the previous example, $\nabla\Phi_{\mathbf{V}}$ is also Nyquist and therefore second and third terms are Nyquist. Since $\Phi_{\mathbf{Z}}$ is given to be Nyquist, the last term is Nyquist too,

making $\nabla\nabla^T H$ Nyquist as desired. Hence,

$$E[\mathbf{e}\mathbf{e}^T/\mathbf{x}] = E[\mathbf{e}\mathbf{e}^T] = E[(\mathbf{z}-\mathbf{v})(\mathbf{z}-\mathbf{v})^T]. \tag{4.55}$$

If the quantization basic cell and the regions of supports of the random vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ are symmetric with respect to the origin, then $E[\mathbf{e}\mathbf{e}^T] = E[\mathbf{z}\mathbf{z}^T + \mathbf{v}\mathbf{v}^T] = E[\mathbf{z}\mathbf{z}^T] + E[\mathbf{z}_1\mathbf{z}_1^T] + E[\mathbf{z}_2\mathbf{z}_2^T]$. Note that the dither in this example satisfies the condition for the first part of the theorem as well, hence the first order moment is also independent of the input. In particular, notice the following special cases:

Assume the quantization basic cell, $\mathcal{P}_0$, is symmetric with respect to the origin.

(i) if both $\mathbf{z}_1$ and $\mathbf{z}_2$ are uniform in $SPD(\mathbf{V})$, then

$$E[\mathbf{e}\mathbf{e}^T/\mathbf{x}] = E[\mathbf{e}\mathbf{e}^T] = \mathbf{G}_D(\mathcal{P}_0, \mathbf{V}) + \frac{1}{6}\mathbf{V}\mathbf{V}^T \tag{4.56}$$

(ii) if both $\mathbf{z}_1$ and $\mathbf{z}_2$ are uniform in $\mathcal{P}_0$, then

$$E[\mathbf{e}\mathbf{e}^T/\mathbf{x}] = E[\mathbf{e}\mathbf{e}^T] = 3\mathbf{G}_D(\mathcal{P}_0, \mathbf{V}) \tag{4.57}$$

Assume we use a dither as in Example 3, which satisfies the first and second order moment independence conditions. Among all such schemes, the minimum total mean square error is achieved by using the lattice quantizer with $\mathcal{P}_0 = VOR(\mathbf{V})$, and a dither vector that is sum of two independent vectors that are uniform in $VOR(\mathbf{V})$ as in the second special case given above. The resulting total mean square error is three times that of the subtractive dithered quantization and that is true for any dimension D. Making use of Theorem 5 on optimum lattices, we have the following result:

Theorem 7. *Let* $\mathbf{V}$ *be the generating matrix of the optimum lattice (i.e., the lattice with minimum* $\sigma_D^2(VOR, \mathbf{V})$*). In subtractive dithering, the minimum total mean square error is achieved by any dither that is Nyquist-*$\mathbf{V}$*. In nonsubtractive dithering, among all dithers as in Example 3, the minimum total mean square error is achieved by the optimal lattice* $\mathbf{V}$*, and by the dither that is the sum of two independent Nyquist-*$\mathbf{V}$ *vectors each of which is uniform in* $VOR(\mathbf{V})$*. The resulting second moment matrices*

are:

$$E[\mathbf{e}\mathbf{e}^T] = \mathbf{G}_D(VOR, \mathbf{V}) = \sigma_D^2(VOR, \mathbf{V})|det\mathbf{V}|^{2/D}\mathbf{I} \quad \text{(subtractive dithering)} \quad (4.58)$$

$$E[\mathbf{e}\mathbf{e}^T] = 3\mathbf{G}_D(VOR, \mathbf{V}) = 3\sigma_D^2(VOR, \mathbf{V})|det\mathbf{V}|^{2/D}\mathbf{I} \quad \text{(nonsubtractive dithering)} \quad (4.59)$$

$\diamondsuit$

Necessary and sufficient condition for total mean square error independence. In Theorem 6, we gave the necessary and sufficient conditions for the first order moment vector and the second order moment matrix of the error to be independent of the input. One can desire to make the total mean square error, $E[\|\mathbf{e}\|^2]$, instead of the second order matrix, $E[\mathbf{e}\mathbf{e}^T]$, independent of the input. The following corollary to Theorem 6 states the necessary and sufficient condition for this weaker requirement:

Corollary 1. In the nonsubtractive quantization scheme of Fig. 4.9, the total mean square error is independent of the input vector, i.e., $E[\|\mathbf{e}\|^2/\mathbf{x}] = E[\|\mathbf{e}\|^2]$, if and only if $Tr(\nabla\nabla^T H(\mathbf{\Omega})$ is Nyquist-$\mathbf{U}$, that is $Tr(\nabla\nabla^T H(\mathbf{Un})) = d\ \delta(\mathbf{n})$. $\diamondsuit$

If the quantization basic cell is symmetric with respect to the origin and if the above condition holds, then $E[\|\mathbf{e}\|^2] = E[\|\mathbf{z}\|^2] + E[\|\mathbf{v}\|^2] = Tr(\mathbf{G}_D(\mathcal{P}_0, \mathbf{V})) + E[\|\mathbf{v}\|^2]$, where $\mathbf{z}$ is defined as in (4.39) and is independent of $\mathbf{v}$.

Proof. From (4.41) in Lemma 1,

$$\begin{aligned} E[\|\mathbf{e}\|^2/\mathbf{x}] &= Tr(\frac{1}{j^2}\sum_{\mathbf{n}} \nabla\nabla^T H(\mathbf{Un}) e^{j\mathbf{x}^T\mathbf{Un}}) \\ &= \frac{1}{j^2}\sum_{\mathbf{n}} Tr(\nabla\nabla^T H(\mathbf{Un})) e^{j\mathbf{x}^T\mathbf{Un}} \end{aligned} \quad (4.60)$$

Hence, by the uniqueness of Fourier series, the necessary and sufficient condition follows. If the condition is satisfied, then $E[\|\mathbf{e}\|^2] = \frac{1}{j^2} Tr(\nabla\nabla^T H(\mathbf{0})) = Tr(E[(\mathbf{z}-\mathbf{v})(\mathbf{z}-\mathbf{v})^T])$, which leads to the result, since $\mathbf{v}$ and $\mathbf{z}$ are independent. ■

Generation of the dither vector for nonsubtractive case: We need a random

vector that is uniform in $VOR(\mathbf{V})$ in the scheme of Example 3 to achieve minimum mean square error. Here is a simple method to generate such a vector: Obtain a dither vector $\mathbf{z}$ that is uniform in $SPD(\mathbf{V})$ using the method given in Sec. 4.2.1. Quantize $\mathbf{z}$ using the lattice quantizer $Q(VOR, \mathbf{V})$. Take the dither vector $\mathbf{v}$ to be the quantization error: $\mathbf{v} = \mathbf{z} - Q(\mathbf{z})$. Then $\mathbf{v}$ is uniform in $VOR(\mathbf{V})$ because of Theorem 1. More generally, one can generate a uniform random vector in any basic cell $\mathcal{P}$ of the lattice $\mathcal{L}(\mathbf{V})$ by replacing the quantizer with $Q(\mathcal{P}, \mathbf{V})$.

Remark. In subtractive dithering, any Nyquist-$\mathbf{V}$ dither produces an error that is independent of the input and uniform in the quantization basic cell. Hence the resulting mean square error is independent of the particular dither used. In nonsubtractive dithering, on the other hand, the total mean square error depends on the dither as well. In particular, the dither should be confined in as small volume as possible in order to obtain the lowest total mean square error.

4.5 Optimum Pre- and Post-filtering For Lattice Quantizers

In traditional scalar quantization schemes where a random process $x(n)$ is uniformly quantized, one assumes that the quantizer noise process $e(n)$ is WSS, white and has a power proportional to the input power. That is, $e(n)$ has a power spectral density $S_{ee}(e^{j\omega}) = c\sigma_x^2$. With these assumptions, one considers the possibility of improvement of the noise level by pre-filtering the input process before quantization and post-filtering it after the quantization with the inverse of the original filter (see Fig. 4.10). It is known

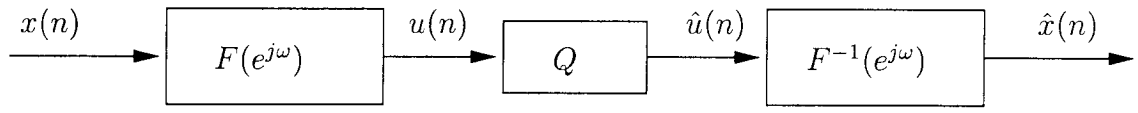


Fig. 4.10: Pre and post-filtering of a scalar process. Q denotes a uniform scalar quantizer. The optimum choice of the filter is the half-whitening solution.

[JN84] that the best pre-filter $F(e^{j\omega})$ is given by

$$|F(e^{j\omega})|^2 = \frac{1}{\sqrt{S_{xx}(e^{j\omega})}} \tag{4.61}$$

and that the phase of $F(e^{j\omega})$ is arbitrary. This is commonly referred as *half-whitening* since the power spectral density of the output of $F(e^{j\omega})$ is $\sqrt{S_{xx}(e^{j\omega})}$, which is flatter than $S_{xx}(e^{j\omega})$ but not completely flat.

The assumptions that lead to the half-whitening solution are valid if the number of levels of the uniform quantizer is very large. However, if one uses a dithered quantizer with proper choice of dither, then the assumptions are not only valid but are precisely true regardless of the bit rate. Hence the half-whitening filter is the optimum filter for a dithered quantizer. After making this elementary observation, we now ask the same question in the lattice vector quantization context: what is the optimum pre-filter matrix $\mathbf{F}(e^{j\omega})$ that produces minimum total mean square error? In this section we proceed to answer this question.

Dithering of WSS vector random processes. Let $\mathbf{x}(n)$ be a WSS vector process with power spectral density matrix $\mathbf{S_{xx}}(e^{j\omega})$. Let $\mathbf{v}(n)$ be a vector process independent of $\mathbf{x}(n)$. Assume we add the two processes together and then quantize the sum at each time instant n with a lattice quantizer $Q(VOR, \mathbf{V})$. After the quantization, we can either subtract the original dither process resulting in subtractive dithering or we can leave it as it is, resulting in nonsubtractive dithering. This is a generalization of Fig. 4.6 and Fig. 4.9, with all the vectors replaced by vector random processes. First consider the subtractive case. It is not difficult to see that, if the dither process is chosen to be iid and Nyquist-$\mathbf{V}$, then the error process $\mathbf{e}(n) = \mathbf{x}(n) + \mathbf{v}(n) - Q(\mathbf{x}(n) + \mathbf{v}(\mathbf{n}))$ will be independent of $\mathbf{x}(n)$ and iid, with uniform distribution in $VOR(\mathbf{V})$. Next, for the nonsubtractive case, if the dither process is chosen to be the sum of two independent random process each of which is iid and uniform in $VOR(\mathbf{V})$, then the second moment of the error vector $\mathbf{e}$ will be independent of $\mathbf{x}(n)$. Assume that we are using the

optimum lattice $\mathcal{L}(\mathbf{V})$ for the given dimension. Then from Theorem 7, we have

$$E[\mathbf{e}(n)\mathbf{e}^T(n+k)] = \sigma_D^2(VOR, \mathbf{V})|det\mathbf{V}|^{2/D}\delta(k)\mathbf{I} \quad \text{(subtractive dithering)} \quad (4.62)$$

$$E[\mathbf{e}(n)\mathbf{e}^T(n+k)] = 3\sigma_D^2(VOR, \mathbf{V})|det\mathbf{V}|^{2/D}\delta(k)\mathbf{I} \quad \text{(nonsubtractive dithering)} \quad (4.63)$$

Pre-filtering of dithered lattice quantization. Assume we filter $\mathbf{x}(n)$ by $\mathbf{F}(z)$ before quantization and by $\mathbf{F}^{-1}(z)$ after the quantization as shown in Fig. 4.11. Let

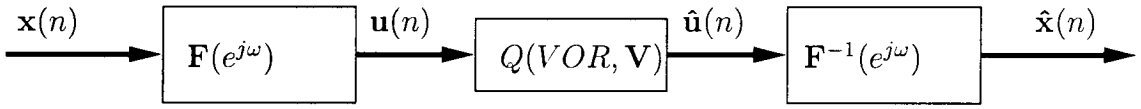


Fig. 4.11: Pre and post-filtering of a vector process in conjunction with a dithered lattice quantizer. The lattice $\mathcal{L}(\mathbf{V})$ is the optimum lattice for its dimension.

$\mathbf{S_{qq}}(e^{j\omega})$ be the power spectral density of the dithered-quantizer noise process $\mathbf{q}(n) = \mathbf{u}(n) - \hat{\mathbf{u}}(n)$. Then, by (4.62) and (4.63), it follows that

$$\mathbf{S_{qq}}(e^{j\omega}) = c\mathbf{I} \quad (4.64)$$

where c depends only on lattice. To be precise, $c = \sigma_D^2(VOR, \mathbf{V})|det\mathbf{V}|^{2/D}$ in subtractive case and $c = 3\sigma_D^2(VOR, \mathbf{V})|det\mathbf{V}|^{2/D}$ in nonsubtractive case.

Assumption about the dependence of c on the input variance. Dithering analysis is valid only if the overflow is avoided. If the total bit rate is constrained to be fixed, then obviously there should be a relation between the unit volume $|det\mathbf{V}|$ of the lattice $\mathcal{L}(\mathbf{V})$ and the statistics of the input. If the bit rate is defined by the logarithm of the total number of codewords, then the support of D-dimensional pdf of the process cannot be infinite. If, on the other hand, the bit rate is defined to be the entropy of the quantized process, then D-dimensional pdf can have infinite support as in the cases of well-known distributions like Gaussian, Laplacian, etc. Without going into the detailed discussion of the rate-distortion analysis of dithered quantizers, we are going to assume that the constant c in (4.64) is proportional to the total variance

of the quantizer input, that is $c = d\sigma_{\mathbf{u}}^2$. Hence (4.64) becomes:

$$\mathbf{S_{qq}}(e^{j\omega}) = d\sigma_{\mathbf{u}}^2\mathbf{I} \tag{4.65}$$

where $\sigma_{\mathbf{u}}^2$ is the total variance of $\mathbf{u}(n)$ in Fig. 4.11.

Theorem 8. *In the scheme of Fig. 4.11, assuming the relation (4.65), the optimum pre-filter matrix that minimizes the total mean square error is given by*

$$\mathbf{F}(e^{j\omega}) = [\mathbf{\Lambda}(e^{j\omega})]^{-1/4}\mathbf{U}(e^{j\omega}) \tag{4.66}$$

where $\mathbf{\Lambda}(e^{j\omega})$ is a diagonal matrix with positive elements, and $\mathbf{U}(e^{j\omega})$ is a paraunitary matrix, i.e., $\mathbf{U}^\dagger(e^{j\omega})\mathbf{U}(e^{j\omega}) = \mathbf{I}, \forall\omega$ [Vai93]. The matrices $\mathbf{U}(e^{j\omega})$ and $\mathbf{\Lambda}(e^{j\omega})$ are related to the power spectral density $\mathbf{S_{xx}}(e^{j\omega})$ of $\mathbf{x}(n)$ as

$$\mathbf{S_{xx}}(e^{j\omega}) = \mathbf{U}^\dagger(e^{j\omega})\mathbf{\Lambda}(e^{j\omega})\mathbf{U}(e^{j\omega}). \tag{4.67}$$

The resulting total mean square error is

$$\sigma_{\mathbf{e}}^2 = d\Big[\int_{-\pi}^{\pi} Tr([\mathbf{\Lambda}(e^{j\omega})]^{1/2})\frac{d\omega}{2\pi}\Big]^2 \tag{4.68}$$

$\diamondsuit$

Proof. Let $\mathbf{S_1}(e^{j\omega}) = \mathbf{U}^\dagger(e^{j\omega})[\mathbf{\Lambda}(e^{j\omega})]^{1/2}\mathbf{U}(e^{j\omega})$, where $\mathbf{U}(e^{j\omega})$ and $\mathbf{\Lambda}(e^{j\omega})$ are as defined in the theorem statement. Then, $\mathbf{S_{xx}}(e^{j\omega}) = \mathbf{S_1}(e^{j\omega})\mathbf{S_1}^\dagger(e^{j\omega})$. Now,

$$\begin{aligned}\mathbf{R_{ee}}(0) = E[\mathbf{e}(n)\mathbf{e}^T(n)] &= \int_{-\pi}^{\pi}\mathbf{F}^{-1}(e^{j\omega})\mathbf{S_{qq}}(e^{j\omega})[\mathbf{F}^{-1}(e^{j\omega})]^\dagger\frac{d\omega}{2\pi} \\ &= d\int_{-\pi}^{\pi}\sigma_{\mathbf{u}}^2\mathbf{F}^{-1}(e^{j\omega})[\mathbf{F}^{-1}(e^{j\omega})]^\dagger\frac{d\omega}{2\pi}\end{aligned} \tag{4.69}$$

$$E[\|\mathbf{e}\|^2] = d\sigma_{\mathbf{u}}^2 Tr\int_{-\pi}^{\pi}\mathbf{F}^{-1}(e^{j\omega})[\mathbf{F}^{-1}(e^{j\omega})]^\dagger\frac{d\omega}{2\pi}$$

$$
\begin{aligned}
&= d\, Tr\int_{-\pi}^{\pi} \mathbf{F}(e^{j\omega})\mathbf{S_{xx}}(e^{j\omega})\mathbf{F}^{\dagger}(e^{j\omega})\frac{d\omega}{2\pi} Tr\int_{-\pi}^{\pi} \mathbf{F}^{-1}(e^{j\omega})[\mathbf{F}^{-1}(e^{j\omega})]^{\dagger}\frac{d\omega}{2\pi} \\
&= d\int_{-\pi}^{\pi} Tr(\mathbf{F}(e^{j\omega})\mathbf{S_{xx}}(e^{j\omega})\mathbf{F}^{\dagger}(e^{j\omega}))\frac{d\omega}{2\pi}\int_{-\pi}^{\pi} Tr(\mathbf{F}^{-1}(e^{j\omega})[\mathbf{F}^{-1}(e^{j\omega})]^{\dagger})\frac{d\omega}{2\pi} \\
&\geq d\Big[\int_{-\pi}^{\pi} \sqrt{Tr(\mathbf{F}(e^{j\omega})\mathbf{S_{xx}}(e^{j\omega})\mathbf{F}^{\dagger}(e^{j\omega}))Tr(\mathbf{F}^{-1}(e^{j\omega})[\mathbf{F}^{-1}(e^{j\omega})]^{\dagger})}\frac{d\omega}{2\pi}\Big]^2 \\
&\geq d\Big[\int_{-\pi}^{\pi} Tr(\mathbf{F}(e^{j\omega})\mathbf{S_1}(e^{j\omega})\mathbf{F}^{-1}(e^{j\omega}))\frac{d\omega}{2\pi}\Big]^2 \\
&= d\Big[\int_{-\pi}^{\pi} Tr(\mathbf{S_1}(e^{j\omega}))\frac{d\omega}{2\pi}\Big]^2 \\
&= d\Big[\int_{-\pi}^{\pi} Tr([\mathbf{\Lambda}(e^{j\omega})]^{1/2})\frac{d\omega}{2\pi}\Big]^2
\end{aligned}
\tag{4.70}
$$

The first inequality is Cauchy-Schwartz inequality for integrals and the equality holds if and only if

$$
Tr(\mathbf{F}(e^{j\omega})\mathbf{S_{xx}}(e^{j\omega})\mathbf{F}^{\dagger}(e^{j\omega})) = k'\, Tr(\mathbf{F}^{-1}(e^{j\omega})[\mathbf{F}^{-1}(e^{j\omega})]^{\dagger}) \text{ for all } \omega \tag{4.71}
$$

The second inequality is another Cauchy-Schwartz inequality, applied to the following inner product space:
$(\mathbf{A}, \mathbf{B}) = Tr(\ \mathbf{B}^{\dagger}\mathbf{A})$, (see for example, [FIS79, p. 360])

$$
|Tr(\mathbf{B}^{\dagger}\mathbf{A})|^2 \leq Tr(\mathbf{A}^{\dagger}\mathbf{A})Tr(\mathbf{B}^{\dagger}\mathbf{B}) \tag{4.72}
$$

with equality if and only if $\mathbf{A} = k\mathbf{B}$. Letting $\mathbf{A} = \mathbf{F}(e^{j\omega})\mathbf{S_1}(e^{j\omega})$ and $\mathbf{B} = [\mathbf{F}^{-1}(e^{j\omega})]^{\dagger}$, we have the second inequality and therefore the equality holds if and only if

$$
\mathbf{F}(e^{j\omega})\mathbf{S_1}(e^{j\omega}) = k\ [\mathbf{F}^{-1}(e^{j\omega})]^{\dagger} \tag{4.73}
$$

or equivalently,

$$
[\mathbf{S_1}(e^{j\omega})]^{-1} = \mathbf{F}^{\dagger}(e^{j\omega})\mathbf{F}(e^{j\omega}) \tag{4.74}
$$

where $\mathbf{S_1}(e^{j\omega})$ is the spectral factor of $\mathbf{S_{xx}}(e^{j\omega})$, i.e., $\mathbf{S_{xx}}(e^{j\omega}) = \mathbf{S_1}(e^{j\omega})\mathbf{S_1}^{\dagger}(e^{j\omega})$. We can choose $k = 1$ as it will not affect the final result. So, $\mathbf{F}(e^{j\omega})$ should be a spectral

factor of the inverse of the spectral factor of the positive definite matrix $\mathbf{S_{xx}}(e^{j\omega})$. Note that (4.71) is satisfied automatically if $\mathbf{F}(e^{j\omega})$ is chosen as in (4.74). The filter defined given by (4.66) satisfies (4.74) as can be verified by direct substitution. Hence it is an optimal filter matrix with the resulting total mean-square error as in (4.68). When the dimension is 1, the solution reduces to the well known half-whitening filter as in (4.61). ■

Comment. The solution (4.66) can be understood in the following way: The optimum

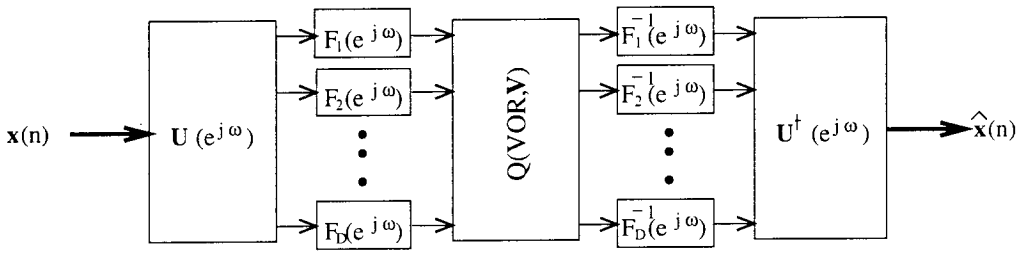


Fig. 4.12: Optimum pre and post-filtering in lattice quantization. $\mathbf{U}(e^{j\omega})$ is the decorrelator filter matrix and the filters $F_1, F_2, ..., F_D$ are the half-whitening filters for their inputs.

$\mathbf{F}(e^{j\omega})$ is the cascade of two systems. The first system, $\mathbf{U}(e^{j\omega})$, which is a paraunitary filter bank, decorrelates the components of the vector process $\mathbf{x}(n)$ (assuming zero-mean for simplicity). The second system, $[\mathbf{\Lambda}(e^{j\omega})]^{-1/4}$, is nothing but half-whitening of each of the decorrelated components! See Fig. 4.12.

Pre-filtered lattice quantization of scalar WSS processes. Assume now that the vector process $\mathbf{x}(n)$ is formed by blocking a WSS random process $x(n)$ [Vai93] (see Fig. 4.13). Then one way to diagonalize the power spectral density is to use a set of

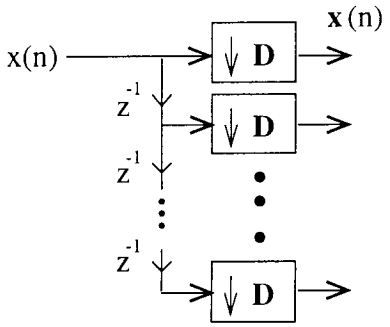


Fig. 4.13: The vector process $\mathbf{x}(n)$, obtained by blocking a scalar WSS process.

ideal filters. Let $\{H_i(e^{j\omega})\}$ be a set of ideal filters that have nonoverlapping frequency supports as shown in Fig. 4.14. Using these filters as in Fig. 4.15, it can be verified that

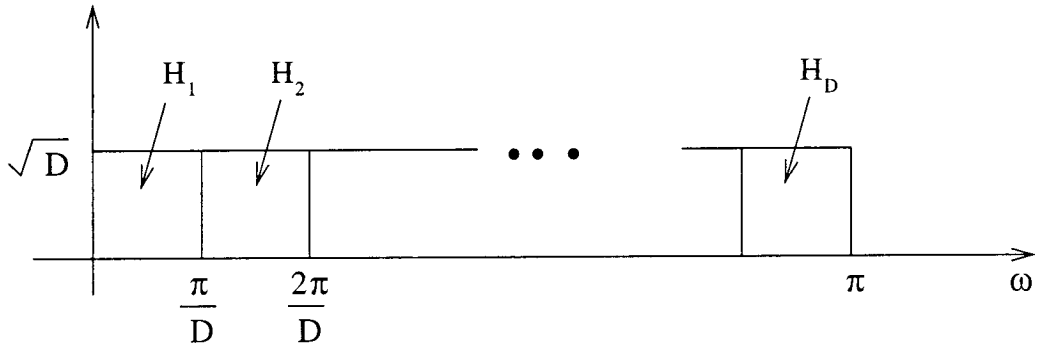


Fig. 4.14: A set of ideal filters to be used as the decorrelating paraunitary system.

the components after the decimation in Fig. 4.15 are uncorrelated. It is not difficult

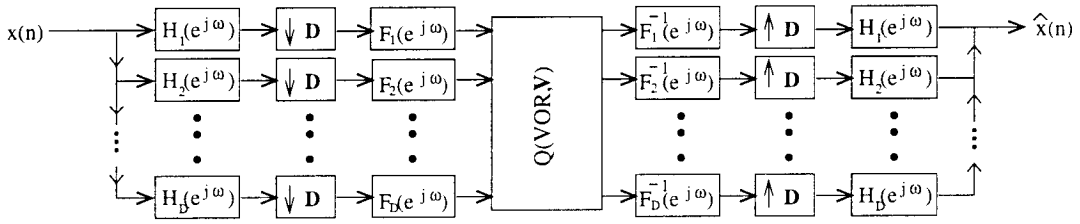


Fig. 4.15: The ideal filter bank of Fig. 4.14 is used as the decorrelating system.

to see that the set of half-whitening pre-filters after the ideal filter bank is equivalent to one half-whitening pre-filter preceding the ideal filter bank. Similarly, the set of corresponding post-filters followed by the ideal filter bank is equivalent to the ideal filter bank followed by one post-filter corresponding to the unblocked output (Fig. 4.16). This system can be redrawn as in Fig. 4.17 using the polyphase representation [Vai93].

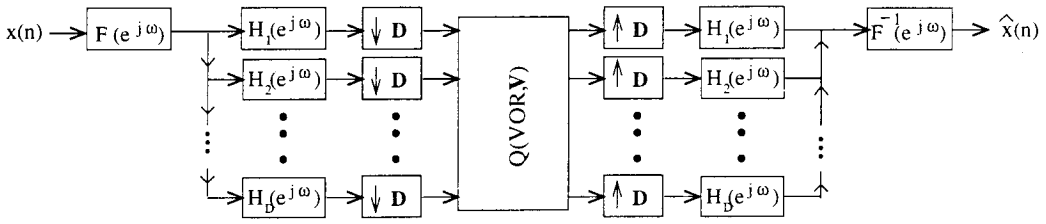


Fig. 4.16: Half-whitening filters reduce to a single half-whitening filter.

By construction, the polyphase matrix $\mathbf{E}(e^{j\omega})$ is paraunitary. Let $\mathbf{r}(n)$ and $\mathbf{u}(n)$ be the

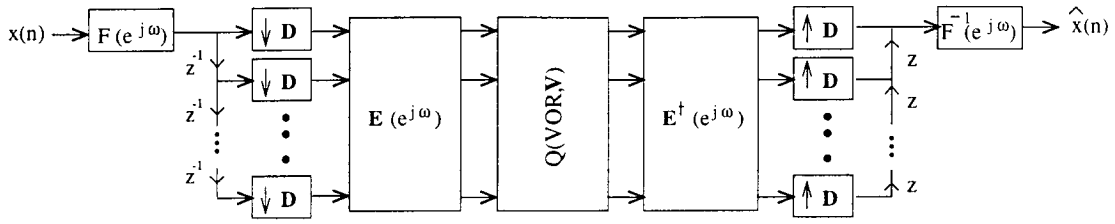


Fig. 4.17: Redrawing the system in Fig. 4.16 using the polyphase decomposition of the ideal filter bank.

input and output of the system $\mathbf{E}(e^{j\omega})$. It can be shown (Appendix C of [Vai93]) that $E[\mathbf{u}^T(n)\mathbf{u}(n)] = E[\mathbf{r}^T(n)\mathbf{r}(n)]$. The quantity $\sigma_{\mathbf{u}}^2$ in (4.65) is $E[\mathbf{u}^T(n)\mathbf{u}(n)]$ with the assumption that the processes have zero mean. Hence $\sigma_{\mathbf{u}}^2$ is unaffected by the choice of $\mathbf{E}(e^{j\omega})$. So we can eliminate $\mathbf{E}(e^{j\omega})$ and $\mathbf{E}^\dagger(e^{j\omega})$ and obtain the simplified form of Fig. 4.18. We have proved:

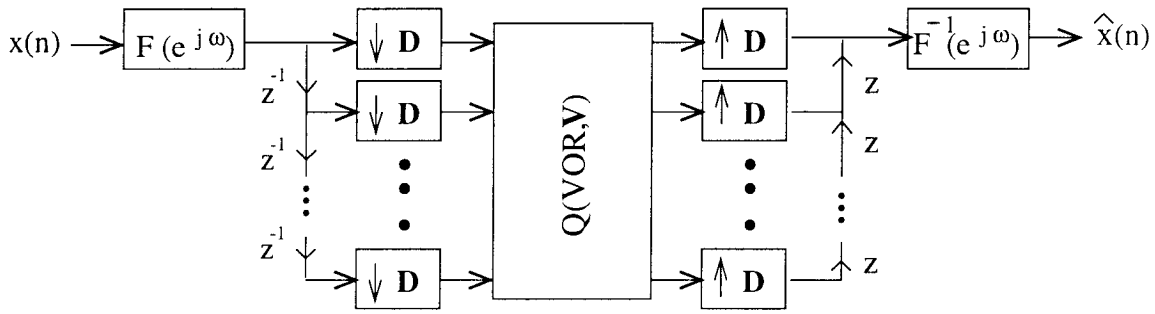


Fig. 4.18: The final simplified form of the system when the input is the blocked version of a scalar WSS process.

Theorem 9. *In the lattice quantization scheme of Fig. 4.11, if the input vector process* $\mathbf{x}(n)$ *is obtained by blocking a WSS scalar process* $x(n)$*, then the optimum pre-filter* $\mathbf{F}(e^{j\omega})$ *is equivalent to the scalar half-whitening filter applied to the input* $x(n)$*, as depicted in Fig. 4.18.* ◇

Relation to the optimum subband coding problem. In subband coding systems, the channels are often quantized with one-dimensional uniform quantizers. Let $u_i(n)$ be the ith subband signal and $q_i(n)$ the corresponding quantization noise. Since each of the channels is quantized separately, the total bit rate is the sum of bit rates of each channel. Let b_i be the rate assigned to the channel i. In subband coding problems, the

following is assumed:

$$\sigma_{q_i}^2 = c\,\sigma_{u_i}^2 2^{-2b_i} \tag{4.75}$$

This assumption is justified when the bit rate is high and the overload effect is negligible [GG92]. The same constant c is assumed for all channels, although in [GG92] it is shown that c depends on the source statistics. For the pre-filtered lattice quantization scheme we assumed (4.65). Since $\sigma_{\mathbf{u}}^2 = \sum_{k=1}^{D} \sigma_{u_k}^2$, this assumption implies

$$\sum_{i=1}^{D} \sigma_{q_i}^2 = d \sum_{i=1}^{D} \sigma_{u_i}^2 \tag{4.76}$$

Compare this with (4.75) which is traditionally used in subband coding with separate subband quantizers. (4.75) yields

$$\sum_{i=1}^{D} \sigma_{q_i}^2 = c \sum_{i=1}^{D} 2^{-2b_i} \sigma_{u_i}^2 \tag{4.77}$$

Thus the set of quantizer noise variances $\{\sigma_{q_i}^2\}$ is assumed to be related to the set of quantizer input variances $\{\sigma_{u_i}^2\}$ by (4.76) in the pre-filtered lattice quantizer, and by (4.77) in the case of traditional subband coding. These two assumptions create significant difference in the formulation and solution of these two problems, which should not, therefore, be compared. In particular, the line of reasoning which allowed us to reduce Fig. 4.17 into the simpler form of Fig. 4.18 will not hold in the traditional subband coding case. As mentioned earlier, the problem of optimizing the pre-filter under the subband coding constraint (4.75) is equivalent to finding the best biorthogonal subband coder for a given input and a fixed number of channels D. The biorthogonal subband coding problem is treated in considerable detail in [VK98a].

4.6 Summary

In this chapter we provided the error analysis of dithered and undithered lattice quantizers. In Sec. 4.2, we analyzed the lattice quantization system. In Sec. 4.3, we saw that, for any input, we can make the quantization error independent from the input

and uniform in the quantization basic cell. We provided some results on the moments of the error and gave a necessary condition for a lattice to have minimum dimensionless second moment. Sec. 4.4 covered nonsubtractive dithering of lattice quantization and we saw that we can make the moments of the error vector independent from the input. We gave one set of dither vectors that can be used in nonsubtractive dithering to achieve the first and second order moment independence conditions. Among them, we outlined how to choose a dither vector that results in minimum total mean square error. We saw that this dither should be a sum of two independent random vectors, each uniform in $VOR(\mathbf{V})$, where $\mathbf{V}$ is the generator matrix of the optimum lattice for its dimension. We emphasized that the requirement to make the total mean square error independent from the input is weaker than the requirement to make the second moment matrix independent from the input. We provided two methods of generating Nyquist-$\mathbf{V}$ vectors, one for the dither that is uniform in $SPD(\mathbf{V})$, the other for the dither that is uniform in $VOR(\mathbf{V})$. The former was sufficient for all purposes in subtractive dithering and the latter was necessary to have minimum mean square error in nonsubtractive dithering. Finally, using the results on optimum lattices from Sec. 4.3, in Sec. 4.5, we addressed the problem of optimum linear pre-filtering of dithered lattice quantizers. With the assumption that the sum of the variances of the noise vector components is proportional to the sum of the variances of the input components, we came up with a general solution. In the special case of blocking one-dimensional WSS processes, we saw that our solution reduces to the scalar half-whitening filter.

APPENDIX A

The definitions of multidimensional Fourier transform, Fourier series and their interrelations are summarized here in a way most suited to our notations. Details can be found in many standard references, for example [DM84].

1. The MD Fourier transform of $f(\mathbf{x})$ is defined as

$$F(\mathbf{\Omega}) = \int f(\mathbf{x}) e^{-j\mathbf{\Omega}^T\mathbf{x}} d\mathbf{x} \tag{4.78}$$

We see that the characteristic function (4.7) is therefore $\Phi_{\mathbf{X}}(\mathbf{\Omega}) = F(-\mathbf{\Omega})$.

2. $f(\mathbf{x})$ is said to be periodic-$\mathbf{V}$, if $f(\mathbf{x}+\mathbf{Vn}) = f(\mathbf{x})$ for every $\mathbf{x} \in R^D$ and $\mathbf{n} \in Z^D$. Let $\mathcal{P}$ be a basic cell with respect to $\mathbf{V}$, and let $\mathbf{U}$ be the matrix generating the reciprocal lattice, that is, $\mathbf{U} = 2\pi\mathbf{V}^{-T}$. Then the Fourier series coefficients of $\mathbf{f}(\mathbf{x})$ are given by

$$c_{\mathbf{k}} = \frac{1}{|det\mathbf{V}|}\int_{\mathcal{P}} f(\mathbf{x})e^{-j\mathbf{x}^T\mathbf{U}\mathbf{k}}d\mathbf{x} \tag{4.79}$$

and the Fourier series representation of $f(\mathbf{x})$ is given by

$$f(\mathbf{x}) = \sum_{\mathbf{k}} c_{\mathbf{k}}e^{j\mathbf{x}^T\mathbf{U}\mathbf{k}} \tag{4.80}$$

3. *Relation between Fourier series and Fourier transform.* Let $F(\mathbf{\Omega})$ be the FT of $f(\mathbf{x})$. Define the periodic-$\mathbf{V}$ function $g(\mathbf{x}) = \sum_{\mathbf{k}} f_{\mathbf{X}}(\mathbf{x} + \mathbf{Vk})$, and let $\{c_{\mathbf{k}}\}$ be its Fourier series as defined above. Then the Fourier series coefficients $\{c_{\mathbf{k}}\}$ are related to the samples of the Fourier transform, taken on the lattice generated by $\mathbf{U}$. More precisely,

$$c_{\mathbf{k}} = \frac{1}{|det\mathbf{V}|}F(\mathbf{U}\mathbf{k}) \tag{4.81}$$

Thus, the periodic function $g(\mathbf{x})$ can be expanded as

$$g(\mathbf{x}) = \frac{1}{|det\mathbf{V}|}\sum_{\mathbf{k}} F(\mathbf{U}\mathbf{k})e^{j\mathbf{x}^T\mathbf{U}\mathbf{k}} \tag{4.82}$$

APPENDIX B

Proof of Fact 1.

$$\begin{aligned}
\mathbf{G}_D(VOR, \mathbf{V}) &= \frac{1}{|det\mathbf{V}|}\int_{VOR(\mathbf{V})} \mathbf{e}\mathbf{e}^T d\mathbf{e} \\
&= \frac{1}{|det\mathbf{V}|}\int_{SPD(\mathbf{V})} \mathbf{e}\mathbf{e}^T d\mathbf{e} \\
&= \mathbf{V}\int_{[-\frac{1}{2},\frac{1}{2})^D} \hat{\mathbf{e}}\hat{\mathbf{e}}^T d\hat{\mathbf{e}}\mathbf{V}^T \\
&= \frac{1}{12}\mathbf{V}\mathbf{V}^T
\end{aligned}$$

$$= \frac{1}{12}|det\mathbf{V}|^{2/D}\mathbf{\Lambda} \tag{4.83}$$

The reason for the second equality is that $VOR(\mathbf{V}) = SPD(\mathbf{V})$ for an orthogonal lattice. The third equality follows by a change of variable $\hat{\mathbf{e}} = \mathbf{V}^{-1}\mathbf{e}$. ■

Proof of Fact 2.

$$\begin{aligned}
\sigma_D^2(\mathcal{P}_0, \mathbf{V}) &\geq \sigma_D^2(VOR, \mathbf{V}) \quad \text{(by definition of } VOR(\mathbf{V})) \\
&= \frac{1}{D|detV|^{2/D}} Tr(\mathbf{G}_D(VOR, \mathbf{V})) \\
&= \frac{1}{12D} Tr(\mathbf{\Lambda}) \quad \text{(by equation (4.21))} \\
&= \frac{1}{12D} \sum_{i=1}^{D} \lambda_i \\
&\geq \frac{1}{12} (\prod_{i=1}^{D} \lambda_i)^{1/D} \quad \text{(arithmetic-geometric mean inequality)} \\
&= \frac{1}{12}
\end{aligned} \tag{4.84}$$

The first inequality can be viewed as an application of the necessary condition for an optimal quantizer: the partition of the space for a given codeword should be the Voronoi partition. It is not difficult to see that no other partitioning can give a better error. Hence, equality holds if and only if $\mathcal{P}_0 = VOR(\mathbf{V})$. The other inequality is an application of arithmetic-geometric mean inequality (abbreviated as AM-GM) [Vai93] to the positive diagonal elements λ_i. Hence, the equality holds if and only if $\lambda_i = c, \forall i$. Finally, because of the definition of $\mathbf{\Lambda}$ in (4.21), $\prod_{i=1}^{D} \lambda_i = 1$, implying $c = 1$. Hence the equality holds if and only if $\mathbf{\Lambda} = \mathbf{I}$. ■

Chapter 5

Concluding Remarks and Future Directions

The main theme of the thesis has been optimal subband coding. We have presented results on the optimality of uniform and nonuniform orthonormal filter banks. We have shown that the problems of optimal orthonormal subband coding and the principal component representation of signals are fundamentally related to each other. Principal component filter banks (PCFB) are those that pack most of the signal energy into any number of retained channels. The special case where the packing is done into one single channel corresponds to the problem of optimal energy compaction filters. We have seen that when we consider the block transforming case, or the ideal subband coding case, there always exist PCFB's and they are optimal for subband coding. The optimality is proven without assuming the traditional high resolution quantizer models. We have extended the results to the nonuniform case, where we have seen that PCFB's can be defined for each of the ordering of the set of subband decimation ratios.

There are still many open problems in optimal subband coding as shown in the review chart in Fig. 1.12. An interesting problem is to find the equivalent of KLT for the nonuniform case. Another interesting problem would be the optimal nonuniform biorthogonal subband coding. In the uniform case, except in the transform coding,

there are advantages of using biorthogonal filter banks. A good review of the current state of knowledge about optimal uniform biorthogonal subband coding can be found in [VK98a].

In the practically interesting case of FIR filter banks, we have seen that the existence of PCFB's cannot be taken for granted as we have demonstrated examples on the contrary. Currently, there is no simple design algorithm that is guaranteed to converge to optimal FIR orthonormal filter banks. Suboptimal solutions do exist, however. One such suboptimal design starts with designing an optimal compaction filter. The completion to a filter bank is done via the canonical factorization of the polyphase vector corresponding to the compaction filter.

We have then presented results on the design of optimal FIR compaction filters. We have developed an analytical method for the special two-channel case. This is the only case where the problems of coding and compaction coincide even with the order constraints on the filters. An FIR PCFB for any input statistics does exist in this case. The reason for such an exceptional result is the fact that one filter of a two-channel orthonormal filter bank determines the other.

We then proposed a very efficient method for designing $M-$channel compaction filters: window method. Although the method is suboptimal, it is very fast as it involves FFT and simple comparison. As the filter order grows, the suboptimality becomes negligible. We compared this technique with the recently introduced technique of linear programming. The latter has the major disadvantage that the complexity grows prohibitively as the filter order increases. We have also proposed multistage design techniques for large order compaction filters. Such techniques result in compaction filters of effective order much higher than the individual filters designed at each stage.

The final chapter has concentrated on the extension of the analysis of uniform quantizers to multiple dimensions. We have provided an exact analysis of lattice quantization noise. We then proposed vector dithering. As in the scalar case, we have considered both the subtractive and nonsubtractive cases. While many of the results are straightforward extension of those in the scalar case, there are some interesting notions in the lattice quantization case that have no counterpart in the scalar case.

One such notion is the selection of the lattice for a given dimension. Best lattices for quantization turn out to be a classical problem for which no solutions are known except for a few small dimensions. We have presented a necessary condition for a lattice to be optimum. We finally have solved the problem of optimal pre- and post-filtering of lattice quantizers.

Bibliography

[ADM95] K. C. Aas, K. A. Duell, and C. T. Mullis. Synthesis of extremal wavelet-generating filters using Gaussian quadrature. *IEEE Trans. on Signal Proc.*, SP-43(5):1045–1057, May 1995.

[AK62] N. I. Ahiezer and M. Krein. *Some questions in the theory of moments.* Translations of Mathematical Monographs, Vol. 2, American Mathematical Society, Providence, Rhode Island, 1962.

[AL91] A. N. Akansu and Y. Liu. On signal decomposition techniques. *Optical Engr.*, 30:912–920, July 1991.

[AM96] K. C. Aas and C. T. Mullis. Minimum mean-squared error transform coding and subband coding. *IEEE Trans. on Inform. Theory*, IT-42(4):1179–1192, July 1996.

[Bau55] F. L. Bauer. A direct iterative process for the Hurwitz-decomposition of a polynomial. *Arch. Elekt. Ubertr., vol. 9, pp. 285-290, June 1955; reviewed by V. Belevitch, IRE Trans. on Circuit Theory*, CT-2:370–371, Dec. 1955.

[BS83] E.S. Barnes and N.J.A. Sloane. The optimal lattice quantizer in three dimensions. *SIAM J. Algebraic Discrete Methods*, 4:30–41, March 1983.

[CLA91] H. Caglar, Y. Liu, and A. N. Akansu. Statistically optimized PR-QMF design. In *SPIE, Visual Comm. and Image Proc. '91: Visual Comm.*, volume 1605, pages 86–94, 1991.

[CLG89] P.A. Chou, T. Lookabaugh, and R.M. Gray. Entropy-constrained vector quantization. *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, 37(1):31–42, Jan. 1989.

[Cox69] H.S.M. Coxeter. *Introduction to geometry.* New York: Wiley, 1969.

[CS82] J.H. Conway and N.J.A. Sloane. Fast quantizing and decoding algorithms for lattice quantizers and codes. *IEEE Trans. on Inform. Theory*, 28:227–232, March 1982.

[CS88] J.H. Conway and N.J. Sloane. *Sphere packings, lattices and groups.* New York: Springer-Verlag, 1988.

[CU82] P. R. Chevillat and G. Ungerboeck. Optimum FIR transmitter and receiver filters for data transmission over band-limited channels. *IEEE Trans. on Comm.*, COM-30(8):1909–1915, Aug. 1982.

[Dav75] P. J. Davis. *Interpolation and approximation.* New York, NY: Dover Publications, Inc., 1975.

[DM84] D.E. Dudgeon and R.M. Mersereau. *Multidimensional digital signal processing.* New Jersey: Prentice Hall, 1984.

[DMS92] P. Delsarte, B. Macq, and D. T. M. Slock. Signal-adapted multiresolution transform for image coding. *IEEE Trans. on Inform. Theory*, IT-38(2):897–904, March 1992.

[DV94] I. Djokovic and P. P. Vaidyanathan. Results on biorthogonal filter banks. *Applied and computational harmonic analysis*, 1:329–343, 1994.

[FIS79] S.H. Friedberg, A.J. Insel, and L. E. Spence. *Linear Algebra.* Englewood Cliffs, NJ: Prentice-Hall, 1979.

[Ger79] A. Gersho. Asymptotically optimal block quantization. *IEEE Trans. Information Theory*, IT-25:373–380, July 1979.

[GG92] A. Gersho and R. M. Gray. *Vector quantization and signal compression.* Boston: Kluwer, 1992.

[GS88] J.D. Gibson and K. Sayood. Lattice quantization. *Advances in Electronics and Electron Physics, (P. Hawkes, Ed.)*, 72(3), 1988.

[GS93] R.M. Gray and T.G. Stockham. Dithered quantizers. *IEEE Trans. Information Theory*, IT-39:805–812, May 1993.

[HJ85] R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge University Press, 1985.

[HS63] Y. Huang and P. M. Schultheiss. Block quantization of correlated gaussian random variables. *IEEE Trans. Comm. Syst*, pages 289–296, Sept. 1963.

[htt] http://www.systems.caltech.edu/kirac/.

[Jai89] A. K. Jain. *Fundamentals of Digital Signal Processing.* Englewood Cliffs, NJ: Prentice Hall, 1989.

[JG93] D.G. Jeong and J.D. Gibson. Uniform and piecewise uniform lattice vector quantization for memoryless gaussian and laplacian sources. *IEEE Trans. Information Theory*, 39:786–804, May 1993.

[JG95] D.G. Jeong and J.D. Gibson. Image coding with uniform and piece-wise uniform vector quantizers. *IEEE Trans. Image Proc.*, 4:140–146, Feb. 1995.

[JLW96] Q. Jin, T. Luo, and K. M. Wong. Optimum filter banks for signal decomposition and its application in adaptive echo cancellation. *IEEE Trans. on Signal Processing*, SP-44(7):1669–1680, July 1996.

[JN84] N. S. Jayant and P. Noll. *Digital coding of waveforms.* Englewood Cliffs, NJ: Prentice Hall, 1984.

[KV95] A. Kıraç and P. P. Vaidyanathan. Dithering in lattice quantization. In *Proc. of the 29th Asilomar Conf. on Signals, Systems, and Computers*, pages 1066–1070, 1995.

[KV96a] A. Kıraç and P. P. Vaidyanathan. Efficient design methods of optimal FIR compaction filters for M-channel FIR subband coders. In *Proc. of the 30th Asilomar Conf. on Signals, Systems, and Computers*, 1996.

[KV96b] A. Kıraç and P. P. Vaidyanathan. Results on lattice vector quantization with dithering. *IEEE Trans. on Circuits and Systems-II*, CAS-43(12):811–826, December 1996.

[KV97a] A. Kıraç and P. P. Vaidyanathan. Analytical method for 2-channel optimum FIR orthonormal filter banks. In *Proc. of IEEE Int. Symp. Circuits and Systems*. Hong Kong, June 1997.

[KV97b] A. Kıraç and P. P. Vaidyanathan. FIR compaction filters : new design methods and properties. In *Proc. of Int. Conf. Acoust. Speech, and Signal Proc.*, pages 2229–2232. Munich, Germany, April 1997.

[KV98a] A. Kıraç and P. P. Vaidyanathan. On existence of FIR principal component filter banks. In *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc.* Seattle, May 1998.

[KV98b] A. Kıraç and P. P. Vaidyanathan. Optimal nonuniform orthonormal filter banks for subband coding and signal representation. In *Proc. of the IEEE Int. Conf. Image Proc.*, 1998.

[KV98c] A. Kıraç and P. P. Vaidyanathan. Optimality of orthonormal transforms for subband coding. In *Proc. of 8th IEEE DSP Workshop*. Utah, August 1998.

[KV98d] A. Kıraç and P. P. Vaidyanathan. Principal component filter banks and optimal orthonormal subband coding : uniform and nonuniform cases. 1998.

[KV98e] A. Kıraç and P. P. Vaidyanathan. Theory and design of optimum FIR compaction filters. *IEEE Trans. on Signal Proc. Special Issue on Theory and Application of Filter Banks and Wavelet Transforms*, 46:903–919, April 1998.

[LWV92] S.P. Lipshitz, R.A. Wannamaker, and J. Vanderkooy. Quantization and dither - a theoretical survey. *J. Audio Eng. Soc.*, 40:355–375, May 1992.

[LZ94] T. Linder and K. Zeger. Asymptotic entropy-constrained performance of tessellating and universal randomized lattice quantization. *IEEE Trans. Information Theory*, IT-40:575–579, March 1994.

[MAKP96] P. Moulin, M. Anitescu, K.O. Kortanek, and F. Potra. Design of signal-adapted FIR paraunitary filter banks. In *Proc. of the IEEE ICASSP-96*, volume 3, pages 1519–1522. Atlanta, May 1996.

[MAKP97] P. Moulin, M. Anitescu, K.O. Kortanek, and F. Potra. The role of linear semi-infinite programming in signal-adapted QMF bank design. *IEEE Trans. on Sign. Proc.*, SP-45(9):2160–2174, Sept. 1997.

[Mal92] H. S. Malvar. *Signal processing with lapped transforms.* Artech House, 1992.

[Mal97] H. S. Malvar. Lapped biorthogonal transforms for transform coding with reduced blocking and ringing artifacts. In *Proc. of Int. Conf. Acoust. Speech, and Signal Proc.*, pages 2421–22424. Munich, Germany, April 1997.

[MM98] P. Moulin and M. K. Mihcak. Theory and design of signal-adapted FIR paraunitary filter banks. *IEEE Trans. on Signal Proc. Special Issue on Theory and Application of Filter Banks and Wavelet Transforms*, 46:920–929, April 1998.

[MO79] A. W. Marshall and I. Olkin. *Inequalities: theory of majorization and its applications.* Academic Press, 1979.

[Mou95] P. Moulin. A new look at signal-adapted QMF bank design. In *Proc. of the IEEE ICASSP-95*, volume 5, pages 1312–1315. Detroit, May 1995.

[NCYM84] Y. Neuvo, D. Cheng-Yu, and S. Mitra. Interpolated finite impulse response filters. *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, ASSP-32(3):563–570, 1984.

[Ngu92] T. Q. Nguyen. A novel spectral factorization method and its application in the design of filter banks and wavelets. In *Proc. IEEE-SP Int. Symp. Time-Freq. Time-Scale Anal.*, pages 303–306. Canada, Oct. 1992.

[Pis73] V. F. Pisarenko. The retrieval of harmonics from a covariance function. *Geophys. J. Roy. Astron. Soc.*, 33:347–366, 1973.

[PMR97] V. Pavlovic, P. Moulin, and K. Ramchandran. An integrated framework for adaptive subband image coding. *IEEE Trans. on Signal Proc.*, 1997.

[RM87] R. A. Roberts and C. T. Mullis. *Digital signal processing.* Addison-Wesley, 1987.

[Rob62] L.G. Roberts. Picture coding using pseudo-random noise. *IRE Trans. Information Theory*, IT-8:145–154, Feb. 1962.

[RVH96] K. Ramchandran, M. Vetterli, and C. Herley. Wavelets, subband coding, and best bases. *Proc. of the IEEE*, 84(4), Apr. 1996.

[Sch64] L. Schuchman. Dither signals and their effect on quantization noise. *IEEE Trans. Comm. Techn.*, Dec. 1964.

[SdS96] V. Silva and L. de Sa. Analytical optimization of CQF filter banks. *IEEE Trans. on Signal Proc.*, SP-44(6):1564–1568, June 1996.

[SG88] Y. Shoham and A. Gersho. Efficient bit allocation for an arbitrary set of quantizers. *IEEE Trans. Acoustics, Speech, and Signal Proc.*, 36(9):1445–1453, Sept. 1988.

[SGR84] K. Sayood, J.D. Gibson, and M.C. Rost. An algorithm for uniform vector quantizer design. *IEEE Trans. Information Theory*, 30:805–814, Nov. 1984.

[Sha93] J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Proc.*, 41(12):3445–3462, Dec. 1993.

[SP96] A. Said and W. A. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. on Circuits and Systems for Video Technology*, 6(3):243–250, June 1996.

[SS77] A.B. Sripad and D.L. Snyder. A necessary and sufficient condition for quantization errors to be uniform and white. *IEEE Trans. Acoust. Speech Signal Proc.*, ASSP-25:442–448, Oct. 1977.

[SS94] M. Shaked and J. G. Shanthikumar. *Stochastic orders and their applications.* Academic Press, 1994.

[SV93] A. Soman and P. P. Vaidyanathan. Coding gain in paraunitary analysis/synthesis systems. *IEEE Trans. on Signal Proc.*, SP-41(5):1824–1835, May 1993.

[TDK95] A. Tirakis, A. Delopoulos, and S. Kollias. Two-dimensional filter bank design for optimal reconstruction using limited subband information. *IEEE Trans. on Image Proc.*, IP-4(8):1160–1165, Aug. 1995.

[TG93] M. K. Tsatsanis and G. B. Giannakis. Time-varying system identification and model validation using wavelets. *IEEE Trans. on Signal Proc.*, SP-41(12):3512–3523, Dec. 1993.

[TG95] M. K. Tsatsanis and G. B. Giannakis. Principal component filter banks for optimal multiresolution analysis. *IEEE Trans. on Signal Proc.*, SP-43(8):1766–1777, Aug. 1995.

[TV98] J. Tuqan and P. P. Vaidyanathan. A state space approach for the design of optimum FIR energy compaction filters. *in preparation*, 1998.

[TZ92] D. Taubman and A. Zakhor. A multi-start algorithm for signal adaptive subband systems. In *Proc. of the IEEE ICASSP-92*, volume 3, pages 213–216. San Francisco, March 1992.

[Uns93a] M. Unser. An extension of the Karhunen-Loeve transform for wavelets and perfect reconstruction filterbanks. In *SPIE, Mathematical Imaging*, volume 2034, pages 45–56, 1993.

[Uns93b] M. Unser. On the optimality of ideal filters for pyramid and wavelet signal approximation. *IEEE Trans. on Signal Proc.*, SP-41(12):3591–3596, Dec. 1993.

[UO95] B. E. Usevitch and M. T. Orchard. Smooth wavelets, transform coding, and markov-1 processes. *IEEE Trans. on Signal Proc.*, SP-43(11):2561–2569, Nov. 1995.

[Vai93] P. P. Vaidyanathan. *Multirate systems and filter banks.* Englewood Cliffs, NJ: Prentice Hall, 1993.

[Vai96] P. P. Vaidyanathan. Theory of optimal orthonormal filter banks. In *Proc. of Int. Conf. Acoust. Speech, and Signal Proc.*, volume 3, pages 1487–1490. Atlanta, May 1996.

[Vai98] P. P. Vaidyanathan. Theory of optimal orthonormal filter banks. *IEEE Trans. on Signal Proc.*, 46(6):1528–1543, June 1998.

[Van92] L. Vandendorpe. CQF filter banks matched to signal statistics. *Signal Proc.*, 29:237–249, 1992.

[VK98a] P. P. Vaidyanathan and A. Kıraç. Results on optimal biorthogonal filter banks. *IEEE Trans. on Circuits and Systems-II, Invited Paper for the Special Issue on Multirate Systems, Filter Banks and Wavelets*, Aug. 1998.

[VK98b] P. P. Vaidyanathan and A. Kıraç. Results on optimum biorthogonal subband coders. In *Proc. IEEE Int. Symp. Circuits and Systems*, pages 154–157. Monterey, CA, June 1998.

[VNDS89] P. P. Vaidyanathan, T. Q. Nguyen, Z. Doganata, and T. Saramaki. Improved technique for design of perfect reconstruction FIR QMF banks with lossless polyphase matrices. *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, 37:1042–1056, 1989.

[Wic94] M. V. Wickerhauser. *Adapted wavelet analysis from theory to software.* IEEE Press, 1994.

[XB98] Bo Xuan and R. H. Bamberger. FIR principal component filter banks. *IEEE Trans. on Signal Proc. Special Issue on Theory and Application of Filter Banks and Wavelet Transforms*, 46:930–940, April 1998.

[ZF92] R. Zamir and M. Feder. On universal qunatization by randomized uniform/lattice quantizers. *IEEE Trans. Information Theory*, IT-38:428–436, Mar. 1992.

[ZF94] R. Zamir and M. Feder. On lattice quantization noise. In *Proc. Data Compression Conf.*, pages 380–389. Snowbird, UT, Mar. 1994.

[ZF95] R. Zamir and M. Feder. Rate-distortion performance in coding bandlimited sources by sampling and dithered quantization. *IEEE Trans. Information Theory*, IT-41, Jan. 1995.

[ZF96] R. Zamir and M. Feder. Information rates of pre/post-filtered dithered quantizers. *IEEE Trans. Information Theory*, 42(5):1340–1353, Sep. 1996.

[Ziv85] J. Ziv. On universal quantization. *IEEE Trans. Information Theory*, IT-31:344–347, May 1985.