# Optimal Source Coding with Signal Transfer Function Constraints

**Milan Stefan Derpich Musa**

***A thesis submitted in partial fulfilment of the requirements for the degree of***

**Doctor of Philosophy**

School of Electrical Engineering
and Computer Science

The University of Newcastle
Callaghan, NSW 2308
Australia

February, 2009

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

Milan S. Derpich

# Acknowledgements

First of all, I would like to thank Professor Graham C. Goodwin for giving me this unique opportunity. Growing as a PhD student under his supervision has been not only a privilege and an honour, but also a truly wonderful experience. I also want to thank Graham for his always friendly attitude towards me, for his patience and guidance, for his permanent support and help and for giving me the space to explore and create. I want to thank my co-supervisor Dr. Daniel E. Quevedo for his careful guidance and orientation, for sharing his knowledge and experience with me, for always finding the way to help me and for his friendship.

I give a special acknowledgement to Dr. Jan Østergaard for his friendly and generous help and for so many things I learned from him. Without these, I wouldn't have been able to address the information theoretic problems of this thesis. Much of my PhD experience has been enriched by the collaboration with my friend and PhD colleague Eduardo I. Silva, whom I want to thank for all those interesting discussions and hours of creative exploration and shared learning. I would also like to thank Dr. Cristian Rojas, Dr. Juan Carlos Agüero, Dr. José De Doná and Dr. María Serón, who so many times helped me overcome theoretical obstacles with their knowledge and good will.

I thank Dianne Piefke and Jayne Disney, for their kind support and thoroughness, which made my stay in Newcastle as well as in other remote lands a hassle free experience.

It was Dr. Mario Salgado who first suggested to me the idea of undertaking a PhD. I want to thank him not only for that, but also for helping it happen and for believing in me.

Finally, I could never overstate my gratitude to my mother, for her love and admirable dedication and discipline, to my father, for his love and his stimulating critical thinking, to my children Ántar and Zoe, for awakening me to happiness, and to my wife and loyal partner Erika, for her love, her support, patience and advice.

# Abstract

This thesis presents results on optimal coding and decoding of discrete-time stochastic signals, in the sense of minimizing a distortion metric subject to a constraint on the bit-rate and on the signal transfer function from source to reconstruction.

The first (preliminary) contribution of this thesis is the introduction of new distortion metric that extends the *mean squared error* (MSE) criterion. We give this extension the name *Weighted-Correlation MSE* (WCMSE), and use it as the distortion metric throughout the thesis. The WCMSE is a *weighted* sum of two components of the MSE: the variance of the error component uncorrelated to the source, on the one hand, and the remainder of the MSE, on the other. The WCMSE can take account of signal transfer function constraints by assigning a larger weight to deviations from a target signal transfer function than to source-uncorrelated distortion.

Within this framework, the second contribution is the solution of a family of feedback quantizer design problems for wide sense stationary sources using an additive noise model for quantization errors. These associated problems consist of finding the frequency response of the filters deployed around a scalar quantizer that minimize the WCMSE for a fixed quantizer *signal-to-(granular)-noise ratio* (SNR). This general structure, which incorporates pre-, post-, and feedback filters, includes as special cases well known source coding schemes such as *pulse coded modulation* (PCM), *Differential Pulse-Coded Modulation* (DPCM), Sigma Delta ($\Sigma\Delta$) converters, and noise-shaping coders. The optimal frequency response of each of the filters in this architecture is found for each possible subset of the remaining filters being given and fixed. These results are then applied to oversampled feedback quantization. In particular, it is shown that, within the linear model used, and for a fixed quantizer SNR, the MSE decays exponentially with oversampling ratio, provided optimal filters are used at each oversampling ratio. If a subtractively dithered quantizer is utilized, then the noise model is exact, and the SNR constraint can be directly related to the bit-rate if entropy coding is used, regardless of the number of quantization levels. On the other hand, in the case of fixed-rate quantization, the SNR is related to the number of quantization levels, and hence to the bit-rate, when overload errors are negligible. It is shown that, for sources with

unbounded support, the latter condition is violated for sufficiently large oversampling ratios. By deriving an upper bound on the contribution of overload errors to the total WCMSE, a lower bound for the decay rate of the WCMSE as a function of the oversampling ratio is found for fixed-rate quantization of sources with finite or infinite support.

The third main contribution of the thesis is the introduction of the *rate-distortion function* (RDF) when WCMSE is the distortion metric, denoted by WCMSE-RDF. We provide a complete characterization for Gaussian sources. The resulting WCMSE-RDF yields, as special cases, Shannon's RDF, as well as the recently introduced *RDF for source-uncorrelated distortions* (RDF-SUD). For cases where only source-uncorrelated distortion is allowed, the RDF-SUD is extended to include the possibility of linear-time invariant feedback between reconstructed signal and coder input. It is also shown that feedback quantization schemes can achieve a bit-rate only $0.254$ bits/sample above this RDF by using the same filters that minimize the reconstruction MSE for a quantizer-SNR constraint.

The fourth main contribution of this thesis is to provide a set of conditions under which knowledge of a realization of the RDF can be used directly to solve encoder-decoder design optimization problems. This result has direct implications in the design of subband coders with feedback, as well as in the design of encoder-decoder pairs for applications such as networked control.

As the fifth main contribution of this thesis, the RDF-SUD is utilized to show that, for Gaussian stationary sources with memory and MSE distortion criterion, an upper bound on the information-theoretic causal RDF can be obtained by means of an iterative numerical procedure, at all rates. This bound is tighter than $0.5$ bits/sample. Moreover, if there exists a realization of the causal RDF in which the reconstruction error is jointly stationary with the source, then the bound obtained coincides with the causal RDF. The iterative procedure proposed here to obtain $\overline{R_c^{it}}(D)$ also yields a characterization of the filters in a scalar feedback quantizer having an operational rate that exceeds the bound by less than $0.254$ bits/sample. This constitutes an upper bound on the optimal performance theoretically attainable by any causal source coder for stationary Gaussian sources under the MSE distortion criterion.

# Contents

# Chapter 1

# Introduction

*Design depends largely on constraints.*
*Charles Ormond Eames, Jr, United States Designer.*

*There exist but two classes of problems in politics: those which solve themselves and those which have no solution.*
*Ramón Barros Luco, former Chilean president (1910-1915).*

*My biggest problem is what to do about all the things I can't do anything about.*
*Ashleigh Brilliant, British Cartoonist.*

## 1.1 Background and Motivation

Many engineering applications require the storage and transmission of signals so that small distortion occurs whilst utilizing a limited number of bits. The achievement of this goal has been one of the fundamental objectives in signal processing research since the beginnings of the "Digital Era" [1–5].

The mathematical characterization of the trade-off between fidelity and data-rate constitutes the essence of what is known as Rate-Distortion Theory [6]. The foundations for this theory were laid by Claude Shannon in [7, 8]. Shannon's *rate-distortion function* (RDF), denoted by $R(D)$, specifies the minimum bit-rate $R$ required for a given amount of distortion $D$ that can be achieved by *any* conceivable source coding system. $R(D)$ has been characterized, to different degrees, for several *probability density functions* (PDFs) and for several distortion metrics [6, 9–11]. By far the best understood case of this rate-distortion trade-off is that which occurs when the source is Gaussian and mean squared error (MSE) is used as the distortion metric. In this case, for a discrete-time Gaussian stationary random source $\{\mathrm{x}(k)\}$

with *power spectral density* (PSD) $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$, the minimum achievable rate for a given distortion $D > 0$ is given by the well known *reverse water-filling equations*[1] [6, 9]

$$R(D) = \frac{1}{2\pi} \int\limits_{\omega: S_{\mathrm{x}}(\mathrm{e}^{j\omega}) > \theta} \frac{1}{2} \log \left( \frac{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}{\theta} \right) d\omega \tag{1.1a}$$

$$D = \frac{1}{2\pi} \int\limits_{-\pi}^{\pi} \min \left\{ \theta, S_{\mathrm{x}}(\mathrm{e}^{j\omega}) \right\} \, d\omega, \tag{1.1b}$$

where $\theta > 0$ is a scalar parameter commonly referred to as the "water level". The relation between rate, distortion and water level can be easily appreciated in the example illustrated in Fig. 1.1-(a). In that

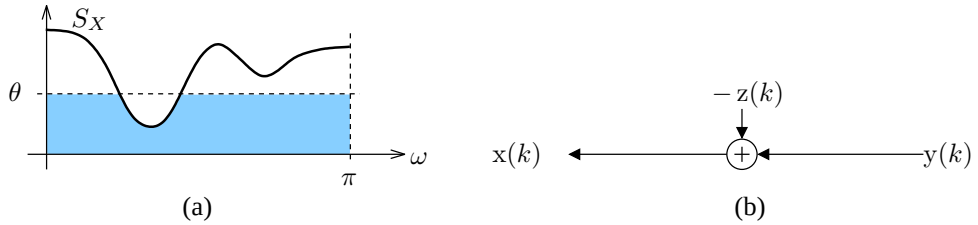



**Figure 1.1:** (a): Graphical representation of the water-filling equations (1.1). The distortion is represented by the colored area under the plot. (b): Backward test-channel realization of $R(D)$.

figure, the distortion $D$ is given by the area under the water level $\theta$ or the plot of $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$, whichever is lower. In turn, only the portion of $S_x(\mathrm{e}^{j\omega})$ standing above the water level contributes to the rate, as can be seen from (1.1a).

Another equally important question in rate-distortion theory is finding a *realization* for the rate distortion function for a source. A realization of a rate-distortion function $R(D)$ corresponds to a probability assignment between source $\mathrm{x}$ and its reconstructed approximation $\mathrm{y}$ such that the distortion is $D$ and the mutual information[2] between $\mathrm{x}$ and $\mathrm{y}$ equals $R(D)$. For a discrete-time Gaussian stationary source with PSD $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$ and using MSE as the distortion metric, $R(D)$ is realized if the optimal reconstruction error

$$\{\mathrm{z}(k)\} \triangleq \{\mathrm{y}(k)\} - \{\mathrm{x}(k)\}$$

is a Gaussian stationary process independent of the output $\{\mathrm{y}(k)\}$, with PSD

$$S_{\mathrm{z}}(\mathrm{e}^{j\omega}) = \min \left\{ \theta, S_{\mathrm{x}}(\mathrm{e}^{j\omega}) \right\}. \tag{1.2}$$

[1] These equations were first derived by Kolmogorov for continuous-time Gaussian sources in [12].

[2] The notion of mutual information is formally introduced in Section 2.3.2 of this thesis.

From the above argument it can be seen that the realization of $R(D)$ can be represented by a *test channel* such as the one shown in Fig. 1.1-(b). Since the additive noise $\{\mathrm{z}(k)\}$ is assumed independent of any other signal *entering* the channel, the arrows point to the left. (Recall that the error must be independent of the output.) Such flow of signals, which may at first sight seem counterintuitive, is an indication of the fact that Shannon's rate-distortion function cannot be realized causally. This implies that, in practice, infinite delay from source to reconstruction would be required to achieve[3] $R(D)$.

Part of the applicability of rate-distortion theory stems from the fact that knowledge of the RDF, for a given source and distortion metric, can be used as a guideline to design rate-distortion efficient *encoder-decoder* (ED) systems, see [13] and the references therein. Indeed, the water-filling equations (1.1) naturally suggest practical coding paradigms such as sub-band coding and transform coding [13, 14]. Moreover, knowledge of the realization of an RDF can, in principle, be utilized as the key to solve optimal source-coder design situations [15]. Unfortunately, this doesn't seem to be the case for most practical ED design problems. In fact, a number of limitations usually arise in practice that preclude the use of the RDF and its realization to aid in the design of ED pairs. As a consequence, not only can the performance of an ED system significantly depart from $R(D)$, but also a designer, if aiming for rate-distortion efficiency, will have to solve an (often difficult) constrained optimization problem. A brief list of these practical limitations includes:

1. the analytical intractability of meaningful *distortion metrics*;
2. *signal transfer function constraints*;
3. *architectural limitations*;
4. *quantization and entropy coding constraints*; and
5. *delay*.

Each of these limitations is briefly discussed below.

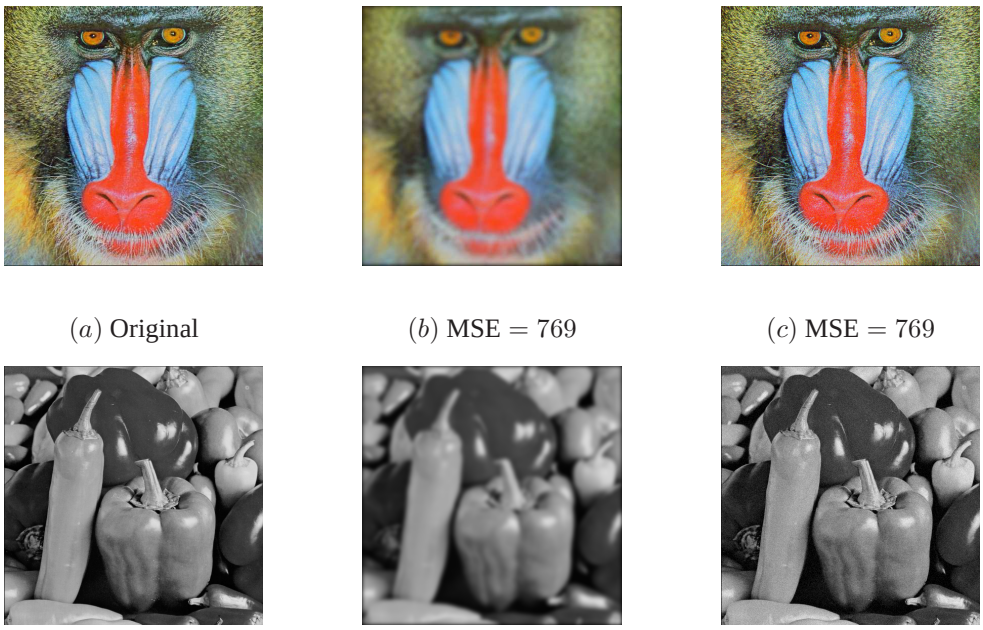
### 1.1.1 Distortion Metrics

It is often the case that the most meaningful distortion metrics for the application of interest make the analytic derivation of the corresponding RDF a formidable, or even impossible, task[4] [19,20]. This is the case of, for example, elaborate distortion metrics based upon perception models of human hearing and

[3]Notice that, in this thesis, *to achieve* means to construct an ED pair that attains an operational bitrate $R(D)$ when the distortion is $D$, which is more restrictive than *to realize* a rate-distortion function.

[4]Except, in some cases, at asymptotically high rates, see, e.g., [16–18].

vision [20–23]. The opposite situation occurs with the MSE distortion metric. MSE is highly amenable to analytical manipulation, but fails to adequately describe perceived distortion in applications such as image processing [23]. An example of this fact is illustrated in Fig. 1.2. In this figure, the perceived degradation in image quality produced by linear distortion (low-pass filtering in images (b) and (e)) is clearly more significant than that due to additive noise uncorrelated to the original picture (images (c) and (f)). Nevertheless, the MSE in (b) is equal to the MSE in (c). The same applies to images (e) and (f).



$(a)$ Original $(b)$ MSE $= 769$ $(c)$ MSE $= 769$

$(d)$ Original $(e)$ MSE $= 36.6$ $(f)$ MSE $= 36.6$

**Figure 1.2:** Comparison between the perceived effect of parallel and uncorrelated distortions. (a), (b): Original images; (b) and (e): Low-pass filtered versions (Gaussian blurring); (c) and (f): Uniformly distributed white noise uncorrelated to the original image added.

### 1.1.2 Signal Transfer Function Constraints

A number of applications impose constraints on the transfer function from source to the reconstructed output. This situation arises, for example, when the output of the decoder is added to the output of one or more other decoders that generate correlated versions of the same source. Typical examples can be found

in sub-band coders [24–28] and in parallel quantization schemes [29–31]. Although, in these cases, the signal transfer functions of all ED pairs could, in principle, be optimized globally, there are situations in which the design must be carried out in a modular fashion. This happens, for example, when the globally optimal design is unknown, or when the other parallel encoder-decoder pairs have been pre-designed and cannot be modified.

Another scenario in which the signal transfer function of an ED pair is important is when the decoded output is fed-back to the encoder input (together with other signals) through an external feedback loop. As an illustration, consider a simple networked control system, as depicted in Fig. 1.3. In this scheme,

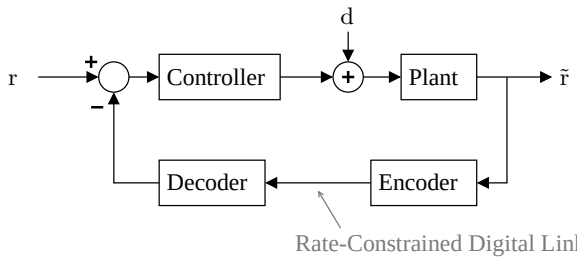



**Figure 1.3:** A simple networked control system. $\mathrm{r}$, $\mathrm{d}$ and $\tilde{\mathrm{r}}$ represent reference, disturbance, and plant output signals, respectively.

let us assume that the reference signal $\mathrm{r}$ and the disturbance signal $\mathrm{d}$ are *wide-sense stationary* (w.s.s.) processes, and that the controller, plant, encoder and decoder are modeled as LTI systems. Suppose that the controller has been designed, in some optimal sense, *without taking into account* the effect of an ED pair in the feedback path. This is indeed the case in many practical situations, either because the feedback path was originally a transparent, noise-less analog link, or because the joint optimal design of controller and encoder-decoder is an open problem [32]. In this situation, if an ED pair is inserted in the loop, as shown in Fig. 1.3, then its associated signal transfer function will affect the dynamic behaviour of the entire closed loop system. This may severely alter dynamic properties such as rise times, settling times, and overshoot. It can also have an impact on the disturbance rejection capabilities of the closed loop control system. More importantly, the open-loop signal transfer function of the ED pair should not have a negative impact on the stability of the closed loop system. This requirement may be particularly restrictive for unstable plants. In this situation, it is often reasonable to design the ED pair so that it exhibits a unit signal transfer function. In this case, the ED pair will not affect the dynamical properties intended in the original closed-loop design.

### 1.1.3 Architectural Limitations

All digital source coding systems are based upon some form of quantization, usually in combination with other blocks, such as filters. In this thesis we restrict attention to cases where all these other blocks are linear. In addition, when the source is an infinite-length random process, we will consider, apart from the quantizer, only linear blocks which are also time-invariant (LTI). Excluding non-linear processing from our analysis leaves aside techniques such as those based on consistent estimates, see. e.g. [33–36], encoding paradigms based on matching pursuit [36, 37], and situations in which all the processing around the quantizer is linear but possibly time-varying [38–41]. It has been shown in [33–36] that non-linear techniques provide improvements in the reconstruction accuracy at the cost of additional complexity, when compared to linear methods. However, encoding schemes based upon linear pre-and post-processing of signal samples around a quantizer are widely used in practice due to their relative computational simplicity.

When all the processing stages around the quantizer are linear, **only three degrees of freedom are available**, namely: (i) to act on the signal before the quantizer (*linear pre-processing*), (ii) to act on the signal after the quantizer (linear *post-processing*), and (iii) to re-inject the output signal (possibly linearly processed) to the input of the quantizer (*linear feedback*). These three degrees of freedom are illustrated in Fig. 1.4. This architecture includes, as special cases, scalar full-band source coding schemes such

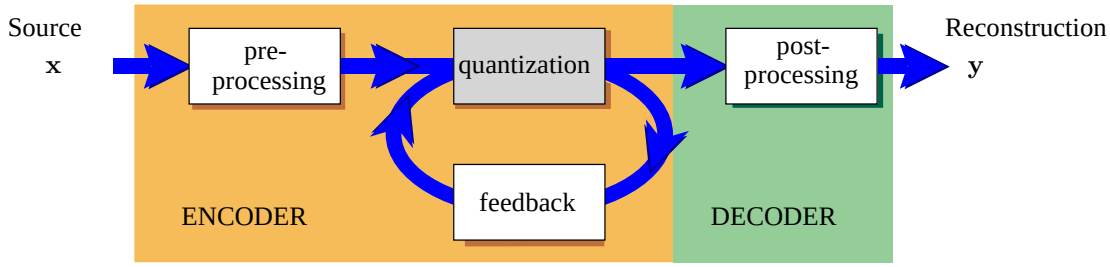



**Figure 1.4:** The three degrees of freedom in any scheme that combines quantization with linear processing blocks: pre-processing, post-processing, and feedback.

as *sigma-delta* ($\Sigma\Delta$) converters [42, 43], *multi stage noise shaping* (MASH) modulators [44, 45], noise shaping quantizers [46–49], delayed-decision or "look-ahead" feedback quantizers [50–54] and DPCM converters [55, 56], as well as subband coders such as transform coders [57, 58] and filter-banks [28, 59–61].

There exist design situations in which one or more of the above degrees of freedom is not available. For example, if an audio encoder is to be designed for standard compact-disc players, the signal processing associated with play-back is fixed, and thus only two degrees of freedom are available: pre-

processing and feedback. As another example, consider an optical sensor with quantized, discrete-time output. In such a device, the physical variable of interest will reach the input of the (internal) quantizer through some transfer function (different from unity), given by the dynamic properties of the transducer within the sensor. This transfer function, which can be seen as linear pre-processing, cannot be modified (unless, of course, internal transducer reconfiguration is viable). This leaves only two degrees of freedom available for the design of encoder and decoder. In relation to any particular situation, we refer to constraints on the degrees of freedom available to the designer of the ED system as **architectural limitations**.

It is natural to expect that any architectural limitation will adversely affect the best achievable performance of an ED system. This raises other questions such as: "How much will the best achievable performance be affected if any of the three degrees of freedom is not available?" "What is the importance of feedback?", "Is feedback always necessary for optimality"? More generally, this motivates the search for the fundamental performance limitations associated with architectural limitations.

### 1.1.4 Quantization and Entropy Coding Constraints

Quantization is the process of mapping continuous amplitude numbers (or vectors) into a finite or countable set of values. Without the use of other processing, vector quantizers are superior in rate-distortion efficiency, when compared to scalar quantizers [62]. Nevertheless, the computational complexity of implementing vector quantization is usually avoided in practice in favor of (simpler) scalar quantizers.

The bit-rate associated with a stand-alone scalar quantizer is given by the number of quantization levels of the latter. More precisely, if the number of quantization levels is $L$, then the binary representation of each quantized output takes $\lceil \log_2(L) \rceil$ bits, where $\lceil \cdot \rceil$ denotes rounding up to the nearest integer. Such combination of quantization and binary-encoding is commonly referred to as *fixed-rate quantization*.

It is well known that *entropy coding* can reduce the average bit-rate (or the total number of bits) required to transmit, or store, the output of a scalar quantizer [63]. Moreover, it has long been recognized that, for memoryless sources, entropy-coded uniform scalar quantization performs very close to Shannon's RDF at all rates [64, 65]. In entropy coded scalar quantization, each possible output that the quantizer can generate is called a *symbol*. An entropy coder maps each of these symbols to (binary) words of different length. For this reason, this combination of quantization and binary-coding is commonly known as *variable-rate quantization*. If the entropy coding mapping is from one symbol to one word (in a sequential fashion) then the word-length (typically in bits) of each word depends on the probability of the corresponding symbol being generated, *conditioned on all previous symbols already generated by the quantizer*. Such an entropy coder will be referred to as *entropy coder with memory*

(ECM).

The computational complexity of implementing ECM is often avoided by using entropy coders that operate *based only upon the marginal probability distribution* of each symbol. The latter corresponds to memory-less entropy coding, since, in this case, past symbols do not participate in the encoding of the current symbol. As expected, the excess bit-rate incurred by ignoring the past in memory-less entropy coding is large if the probabilistic dependence between consecutive symbols is strong, and zero if consecutive symbols are statistically independent, i.e., if the quantizer outputs a memory-less sequence. By using either linear prediction, as in DPCM converters, or narrow-band analysis filters, as in sub-band coders, the memory of the output of the quantizer can be reduced. This can mitigate, and sometimes eliminate, the performance loss of using memory-less entropy coding instead of entropy coding with memory. However, in terms of the general architecture shown in Fig. 1.4, this requires the use of adequate pre-processing and/or feedback. It also requires the freedom to design a matched post-processing stage (e.g., a "colouring" post-filter in the case of predictive quantization, or a synthesis filter bank, in the case of sub-band coding). Thus, when using memory-less entropy coding, the non-availability of any of the design degrees of freedom will have a more adverse effect on the rate-distortion performance of the system, than if an entropy coder with memory could be used.

When encoding a band-limited continuous time source, there existsituations in which increasing the sampling rate is preferable (or less expensive) than increasing the number of levels used for quantization. This is the case in, for example, digital audio [66], [43, Section 1.1]. The practice of sampling a continuous-time source above its Nyquist rate is known as *oversampling*. Notice that oversampling can also be applied to a discrete-time band-limited signal, by creating interpolated samples between the original ones. In both the continuous-time and discrete-time cases, the effect of oversampling is a shrinkage of the support of the spectrum associated with the samples. Oversampling, which can be seen as an increase in time resolution, makes it possible to compensate for poor amplitude resolution, i.e., for coarse quantization. This was recognized early by Bennet in 1948 [67]. However, in achieving this MSE reduction, it is crucial to place appropriate filters around the quantizer [45, 56]. Thus, in oversampled quantization, the optimal use of the three degrees of freedom shown in Fig. 1.4 plays a fundamental role.

### 1.1.5 Delay

It is known that for Gaussian sources with MSE as the distortion metric, any realization of the $R(D)$ function requires infinite (in practice, very large) delay. In particular, the forward-channel realization of $R(D)$ would require the use of non-causal (non-realizable) filters (see, e.g., [6, Section 4.5]). This limits the usability of the RDF (and its associated realizations) for the design of optimal ED pairs subject to

delay constraints. Delay constraints are present, for example, in real-time speech communications and networked control applications. In the latter case, a delay between the source and its reconstruction is highly undesirable, since it severely affects the dynamic properties and the noise rejection capabilities of the closed loop system. Indeed, too much delay can easily render the closed loop system unstable, with catastrophic consequences [68, 69].

The causal rate-distortion function has been characterized only for memoryless sources or in the limit as the rate tends to infinity, see, e.g., [70] and the references therein. For sources with memory, and at medium or low rates, little is known. The solution to the causal rate-distortion problem could be helpful in the design of rate-distortion efficient, causal source coders and decoders.

In addition to the causality of the encoder-decoder pair, a delay constraint may also produce further performance degradation, if the system is subject to additional limitations. For example, if the digital communications link between encoder and decoder is of limited instantaneous capacity, then the use of entropy coding (either ECM or MEC), with average bit-rate close to that capacity, will induce time varying delays. If these delays are not tolerated by the application, then fixed-rate quantization must be used. By using non-uniform quantizers, such as the Max-Lloyd quantizer, the bit-rate of fixed-rate quantization can be very close to that of a uniform quantizer with entropy coding (which has variable-rate), see, e.g., [71]. However, if only "off the shelf" uniform quantizers are available, then the fixed-rate quantization constraint imposed by low delay requirements will entail a performance loss additional to that already inflicted by the need to use a causal ED pair.

In all of the above situations, the design of optimal encoders and decoders cannot benefit from knowledge of the corresponding RDF and its realizations, *unless the RDF has been derived taking account of the constraints associated with the design problem*. As a consequence of this, the search for optimal ED pairs for constrained scenarios has to be undertaken, in some sense, "from scratch".

This is, indeed, the central theme in this thesis: the design of optimal ED pairs under the above limitations, by either: i) following a "bottom-up", constrained optimization approach, or ii) finding, when possible, the RDF of the underlying problem and then following a more expedite, "top-bottom" approach.

## 1.2 Previous Related Work

### 1.2.1 MSE Extensions

It is possible to distinguish two philosophies in the solutions that have been proposed in the existing literature for narrowing the gap between sources and measures for which an RDF can be found and those for which it cannot. The first approach is to approximate the application-meaningful (but non tractable) distortion metric by a simplified version more suitable for analysis and optimization, as done, e.g., in [72, 73]. The second approach is to extend RDF-"friendly" distortion metrics to better represent the impact of reconstruction errors in a wider range of applications. For example, the extension of MSE to a frequency-weighted MSE (FWMSE), for which the RDF for Gaussian sources has been derived [74], has found greater acceptance than plain MSE in areas such as audio quantization [46, 75] and image halftoning [76, 77]. Nevertheless, frequency-weighted MSE still fails to adequately measure, for example, the type of perceptual differences that were illustrated earlier in Fig. 1.2.

### 1.2.2 Brief Review of Source Coding Paradigms

For stochastic sources, the problem of optimal design of the linear processing blocks in the architecture shown in Fig. 1.4 has been solved only for MSE, and under certain constraints and assumptions. These results are reviewed below for two important source coding paradigms associated with the scheme depicted in Fig. 1.4.

**Full-Band Coders:**

For a w.s.s. **scalar process** source with a single **scalar quantizer**, the system in Fig. 1.4 can always be re configured to either of the structures depicted in Figs. 1.5 and 1.6. Both are typical schemes that can be used to describe sigma-delta ($\Sigma\Delta$) converters [42], noise shaping quantizers [46] and DPCM converters [55]. In these figures, the blocks $P$, $A$, $H$, $B$, and $F$ are *linear time-invariant* (LTI) filters,

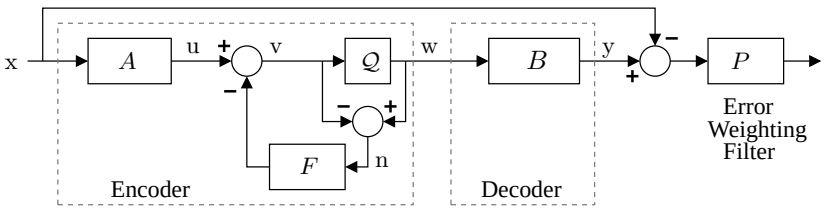



**Figure 1.5:** General feedback quantization system.

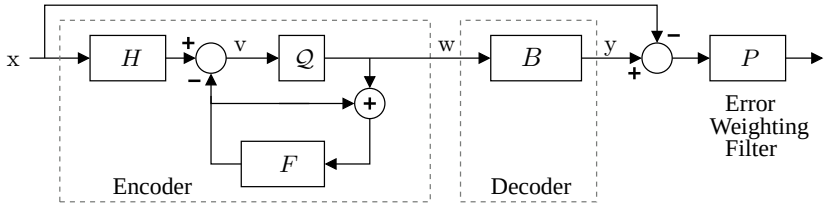



**Figure 1.6:** General feedback quantization system, alternative configuration (equivalent to the one in Fig. 1.5 if and only if $H = A/(1 - F)$).

and $\mathcal{Q}$ represents a scalar quantizer. The feedback filter $F$ needs to be strictly causal (i.e., it must have a delay of at least one sample) for the closed loop to be well defined (see, e.g., [78, Chap. 4]). The block $P$ is a frequency-weighting filter, accounting for the different perceptual impact that reconstruction errors may have at each frequency.

From the viewpoint of the architectural limitations discussed in Section 1.1.3, the system in Fig. 1.6 differs from the one in Fig. 1.5 in that the former does not require being able to measure the signal that enters the quantizer. This is compatible with the architecture limitation in which the pre-processing is given and fixed. By contrast, the configuration shown in Fig. 1.5, in which one can both inject a signal prior to $\mathcal{Q}$ and measure the result, implicitly allows one to arbitrarily modify the pre-processing.

The analysis of the associated feedback system is commonly simplified by modeling the quantization error,

$$\mathrm{n} \triangleq \mathrm{w} - \mathrm{v},$$

as white and uncorrelated with the source $\mathrm{x}$ [55, 79–83]. Hereafter, we will refer to this simplification as the *Linear Model*, to be formally defined later in Section 3.2.2. It must be noted that this model is actually exact if uniform quantization with dither (either subtractive [84] or non-subtractive [85]) is used. It can also serve as a useful approximation in other cases [25, 42, 55, 86].

In the Linear Model, the constraint on the bit-rate is usually expressed as a constraint on the SNR of $\mathcal{Q}$, i.e., the ratio between the variances of $\mathrm{v}$ and $\mathrm{n} = \mathrm{w} - \mathrm{v}$ in Figs. 1.5 and 1.6, see, e.g. [55, 79–83]. This ratio is denoted by

$$\gamma \triangleq \frac{\sigma_{\mathrm{v}}^2}{\sigma_{\mathrm{n}}^2}. \tag{1.3}$$

Under these assumptions, the design of a rate-distortion efficient full-band coder can be posed as the minimization of the frequency weighted MSE for a given and fixed value of $\gamma$, over all filters $A$, $B$, $F$ (of un-restricted order) satisfying the given architectural constraints. Table 1.1 lists several possible optimization problems that can be obtained assuming different combinations of filters as being given and

fixed. We can see from Table 1.1 that some, but not all cases, have been studied earlier. However, these

**Table 1.1:** Architecture-constrained optimization problems. All cases but Case 7 are associated to the system in Fig. 1.5. Case 7 is associated to the configuration shown in Fig. 1.6. $\gamma$ denotes the SNR of $\mathcal{Q}$.

| Case | Optimization Problem | | Existing Results (for MSE Only) | Note |
|---|---|---|---|---|
| | Given $\gamma$, $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$ and | Find the MSE Optimal | | |
| 1 | $A$, $F$ | $B$ | Solution is the standard Wiener filter. | - |
| 2 | $B$, $F$ | $A$ | Solution is the standard Wiener filter | - |
| 3 | $F$, $AB = W$ | $A$ and $B$ | Case $W = 1$, $F = 0$ solved by Noll in [81]. | * |
| 4 | $F$ | $A$ and $B$ | Case $F = 0$ solved by Tuqan and Vaidyanathan in [87]. | ** |
| 5 | $A$, $B$ | $F$ | Results unavailable. The case $A = B = 1$ is a noise-shaping quantizer | *** |
| 6 | $B$ | $A$ and $F$ | Results unavailable | - |
| 7 | $H$ | $B$ and $F$ | Results unavailable | - |
| 8 | - | $A$, $B$ and $F$ | Solved by Zamir, Kochman and Erez in [15]. | ** |

* The optimal system has "half-whitening" pre-and post-filters

** For the case $\gamma \gg \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|F(\mathrm{e}^{j\omega})\right|^2 d\omega$, this problem was first solved in [81].

*** For the case $\gamma \gg \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|F(\mathrm{e}^{j\omega})\right|^2 d\omega$, this problem was first solved in [81]. The solution was then re-derived in [88] under the same assumption.

cases consider the MSE criterion only. In this thesis, we will provide a solution to all the problems in this table as well as extend all of the results to a distortion criterion we propose (which receives here the name *weighted correlation* MSE).

The optimization problem in the last row of Table 1.1 is of particular interest for this thesis. It is a clear illustration of the fact that knowledge of a realization of the underlying RDF of a problem can serve to solve the optimal design problem for an ED pair. This optimization problem had remained open for decades. The following is a brief overview of its history:

To the best of the author's knowledge, the first paper to look for the MSE optimal filters $A$, $B$ and $F$,

for a given $\gamma$, was written by Kimme and Kuo [89] in 1963. In that paper, closed form expressions were derived for optimal frequency responses of the filters $A$ and $B$ as a function of $F$. These expression were exact only for the cases in which $MSE < \min_\omega S_{\mathrm{x}}(\mathrm{e}^{j\omega})$. The optimal solution for $F$ had to be found iteratively over the space of all causal and stable filters. In 1969, Brainard and Candy [80] proposed the design of the corresponding filters combining some of the results in [89] together with heuristic criteria for optimal quantization of television signals. One year later, Noll [81] presented simple analytical expressions for the optimal filters. These expressions were obtained under the simplifying assumption of negligible quantization fed back error. Noll showed that, under this assumption, the optimal filters must necessarily whiten the input signal prior to quantization and, at the same time, yield quantization errors whose PSD becomes white after the error weighting filter. In [82], Atal and Schroeder study the problem of optimal filters for noise-shaping-DPCM converters, focusing on the encoding of speech signals. These authors propose a refined method for the design of the prediction filters, matched to the characteristics of human speech. However, the design of the filter that determines the noise-shaping characteristics of the converter is based on heuristics. After a surprising gap of approximately twenty years without further attempts to solve this problem analytically, an important new insight came with the work of Guleryuz and Orchard in 2001 [90]. As in [89], the analysis in [90] yields analytical expressions for two of the three optimal filters using a Lagrangian approach. Here too, one of the filters has to be found by numerical iteration over the space of all admissible filters. However, unlike [89], these expressions are exact for all distortion values, i.e., for all bit-rate regimes. More importantly, [90] seems to have been the first paper to study the rate-distortion performance of DPCM at low bit-rates, suggesting that scalar quantization with feedback is (nearly) optimal, not only at high rates (as recognized in, e.g., [55, 59]), but at low bit rates as well. A fully analytical solution to this optimization problem finally appeared in recent work by Zamir, Kochman and Erez [15]. Of key importance is the fact that the solution in [15] was *not* derived by solving a constrained optimization problem. Instead, the authors in [15] start from *knowledge of the forward channel realization of Shannon's Rate-Distortion function for Gaussian sources with memory*. Working from this, and following a *deductive* method based on mutual-information equalities, optimal performance, using an entropy coded dithered quantizer, is shown to be $0.254$ bits per sample above Shannon's $R(D)$, *at all rates*. In [15], the optimal filters are not explicitly characterized, but closed form expressions can be obtained from other results in the paper after some additional work. This example illustrates one of the key ideas to be developed in this thesis: *To determine how and when the design of optimal ED pairs can be solved directly from knowledge of a realization of the underlying rate-distortion function.*

It is also of practical importance to characterize the best attainable performance of the scalar feedback

quantization system in Fig. 1.5 as a function of the **oversampling** ratio. As mentioned in Section 1.1.4, oversampling (i.e., sampling a band-limited continuous-time signal at a frequency above its Nyquist rate) allows one to achieve a smaller MSE error for a given, fixed number of quantization levels. For instance, the MSE of simple scalar quantization (without feedback) is known to decrease as $\lambda^{-1}$, where $\lambda$ is the *oversampling ratio*, given by

$$\lambda \triangleq \frac{\text{Sampling Frequency}}{\text{Nyquist Frequency}},$$

see [67]. The latter result has recently been extended to general redundant expansions in [91]. In turn, it has been shown in [56] that feedback scalar quantizers can attain an MSE that is $\mathcal{O}(\lambda^{-2(m+1)})$ as $\lambda \to \infty$, where $m$ is the order of the feedback filter. (See also recent work in [38–40, 92, 93]). From a rate-distortion viewpoint, the inverse polynomial error decay of this error estimate is "too slow" to compensate for the increase in the overall bit-rate due to oversampling (which is proportional to $\lambda$). To be more precise, let us consider a scalar quantizer with $N = 2^b$ quantization levels, where $b$ denotes the quantization resolution in bits per sample. If the additional bit-rate caused by oversampling were to be utilized instead to increase $N$, the MSE would decay as $\mathcal{O}(2^{-2b\lambda})$, i.e., exponentially[5].

A faster decay rate of the MSE of oversampled FQ with $\lambda$ can be achieved by selecting a different feedback filter (of possibly different order) for each oversampling ratio. An example of such a family (of 1-bit $\Sigma\Delta$ converters) was given in [95]. Here, for uniformly bounded inputs, the continuous-time reconstruction error can be uniformly bounded by $\lambda^{-\rho \log \lambda}$, where $\rho > 0$ is independent of $\lambda$. This bound guarantees an MSE that decays with $\lambda$ as $\mathcal{O}(\lambda^{-2\rho \log \lambda})$, which is faster than any inverse polynomial, but still far from exponential. Based on this result, the family of 1-bit $\Sigma\Delta$ converters reported in [96] achieve an MSE that is $\mathcal{O}(2^{-0.14\lambda})$, i.e., exponentially decaying with increasing $\lambda$. Note that the results in [95] and [96] were obtained using an exact, deterministic model for quantization. The author is not aware of results on exponential error decay with oversampling ratio in feedback converters having a multi-bit scalar quantizer or dealing with unbounded support sources.[6]

**Subband Coders:**

The case of the system in Fig. 1.4 in which a w.s.s. source is decomposed into different bands and then quantization is carried out using independent and parallel quantizers corresponds to the typical setting in

[5]Strictly speaking, this has been shown to hold only for signals whose PDFs have finite support. Indeed, it has been shown that for several infinite support source PDFs, the MSE of uniform quantization decreases asymptotically with $b$ not faster than $(\ln 2)^{2/a}\, b^{\frac{2}{a}} 2^{-2b}$, where $a > 0$ is a constant independent of $b$, see [94].

[6]There exist results showing that an exponential error decay with increasing oversampling ratio can be achieved when the quantization threshold crossing *instants* associated with a *continuous time* source are encoded [97–99]. This falls outside of the "first sample and then quantize" paradigm in which this thesis lies.

**filter-banks** (FBs) and *sub-band coding* (SBC) [59]. A typical subband coder is shown in Fig. 1.7. In that figure, the pre-processing stage takes the form of a bank of $M$ analysis filters $H_i(z)$ (analysis filter bank) followed by decimation (down-sampling)[7]. The latter process is represented by the blocks $\downarrow M$, as shown in Fig. 1.7. Each subband signal $\mathrm{u}_i$ is quantized using a separate scalar quantizer, labeled $\mathcal{Q}$.

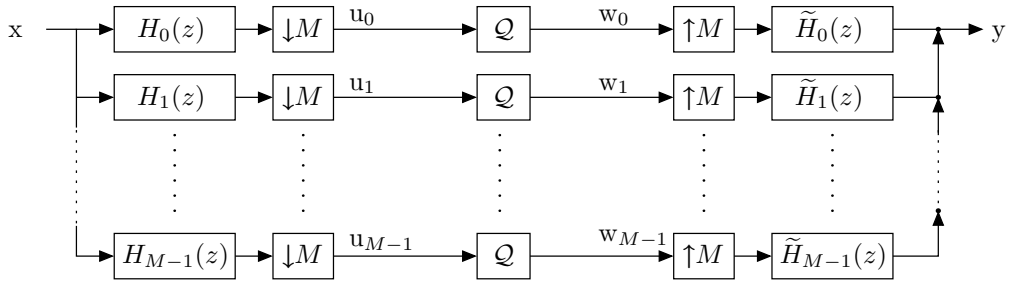



**Figure 1.7:** $M$-channel subband coder with analysis filters $H_i(z)$ and synthesis filters $\widetilde{H}_i(z)$.

The output of each quantizer is then up-sampled $M$-times (zero padding). This operation is represented by the blocks $\uparrow M$ in Fig. 1.7. The output is then fed to the corresponding synthesis filter $\widetilde{H}_i(z)$. The down-sampling allows the inner section of the filter bank to operate at $1/M$ times the sampling rate of the input sequence $\{\mathrm{x}(k)\}$. As a consequence, the total bit-rate is given by the average of the bit-rates associated with each quantizer. Another consequence of decimation is the introduction of a *delay*, of at least $M$ samples, between the source and its reconstruction.

Traditionally, the focus in the subband coding literature has been on *perfect reconstruction* (PR) filter-banks, i.e., on filter-banks where quantization is the only source of reconstruction error (see, e.g., [59, 61, 100–105]). When quantization errors are uncorrelated with the source, the PR condition is a special case of the signal transfer function constraints discussed in Section 1.1.2. However, the motivation for PR in the filter bank literature seems to originate from the search for aliasing-free analysis/synthesis banks in the absence of quantization, rather than being a response to practical situations where a unit signal transfer function could be beneficial. (The papers [106, 107] are an exception.) The study of filter banks that do not satisfy the PR property began a few years later [86, 108–113]. Non-PR filter banks sacrifice the PR property in exchange for achieving lower MSE. However, their superior performance can be critically dependent on accurate knowledge of the statistics of the source [103].

For the case of the *Perfect Reconstruction* constraint, it has been shown by Moulin, Anitescu and Ramchandran that optimal FBs are, in general, biorthogonal [86, Corollary 3.2]. This fact had already

[7] In the equivalent, but computationally more efficient, polyphase representation, the decimation (preceded by different delays) takes place before a (modified) analysis filter bank, see, e.g., [59].

been suggested by the results reported by Aase and Ramstad in [24]. The work in [86] also shows that an optimal PR filter bank can always be constructed as a cascade of a paraunitary *principal component filter bank* (PCFB) followed by a set of pre- and post-filters placed around the quantizer in each subband. The paper [86] additionally gives analytic expressions for the optimal analysis/synthesis FBs, and provides an iterative method to find the optimal bit-allocation.

For the *Non-Perfect Reconstruction* case, expressions for the optimal synthesis FB, for a given analysis FB and a given bit-allocation, have been derived in [114] and [108]. The latter paper also proposes an iterative method for joint design of analysis/synthesis FBs and associated bit-allocation. In a more recent paper [113], Mihçak, Moulin, Anitescu and Ramchandran derive expressions for the optimal analysis/synthesis FBs for a given and fixed bit-allocation. They also propose an iterative algorithm for the computation of globally optimal filters and bit-allocation. It is also shown in [113] that, as in the PR case, an MSE optimal, non-PR FB can always be constructed as a cascade of a paraunitary FB followed by a set of pre- and post-filters placed around the quantizer in each subband. Nevertheless, in general, for the non-PR case, the FB in the first stage of the cascade system, need not be a principal-component filter bank [113, Remark 4].

*Feedback*, i.e., the third degree of freedom in the general architecture depicted in Fig. 1.4, has received relatively scarce attention in the subband coding literature. The use of feedback in subband coding first appeared with the use of DPCM converters, instead of plain scalar quantizers (PCM converters), to quantize each subband signal more efficiently (see, e.g., [115]). Feedback in subband coding has been shown to be beneficial (in the sense of improving rate-distortion performance) in other situations as well. For example, in [61], Bölcskei and Hlawatsch show that feedback is effective in reducing reconstruction MSE in oversampled filter banks. Fisher proved in [116] that the rate of a standard *quadrature mirror filter bank* (QMFB), without feedback, is strictly above the rate-distortion function, except for special cases of the PSD of the source. Then Wong showed in [117] that the use of cross-band prediction (a special case of feedback in the scheme of Fig. 1.4) allows a QMFB to achieve asymptotically optimal rate-distortion performance at high rates. In a recent paper by Makur and Arunkumar [28], the use of feedback is also shown to improve rate-distortion efficiency in biorthogonal subband coders by reducing (and in some cases eliminating) what is known as *quantization noise amplification*. Quantization noise amplification is defined as the ratio between quantization noise variance in the reconstructed signal and the average of the variances of quantization noises introduced in each subband. This ratio is unity for paraunitary filter banks (where there is no quantization noise amplification), but is greater than unity in all perfect reconstruction biorthogonal subband coders [28]. Nevertheless, to the best of the author's knowledge, the problem of *jointly* designing optimal filters and bit allocation for subband coders with

feedback has not been solved. Hence, the optimal achievable performance of a subband coder with three degrees of freedom remains an open problem, except in the limit as the rate, or the number of subbands, tends to infinity.

### 1.2.3 Related Existing Results on Causal and Delay-Free Source Coding

It is known that full-band source coders, such as PCM, DPCM and $\Sigma\Delta$ converters, do not introduce delay between source and reconstructed signal, as long as all the pre- and post-filters do not introduce delay, see Fig. 1.5. However, it is not known how to design an optimal scalar feedback quantizer satisfying a finite- (or zero-) delay constraint. As mentioned in Section 1.2.2, the bit-rate of the predictive converters found in [15], which use subtractive dither and entropy coding, is only $0.254$ bits per sample above Shannon's RDF, for Gaussian stationary sources. Unfortunately, the converters described in [15] require the use of a non-causal pre- or post-filter. In practice, these would need to be approximated by filters which introduce a (possibly) long delay. On the other hand, all the filters in the optimal perfect reconstruction feedback quantizers obtained by the author (and colleagues) in [118] are causal. Therefore, these converters can be considered the best zero-delay source coders for w.s.s. sources described to date. However, being PR converters, it is clear that these ED pairs are still sub-optimal, within the class of zero-delay source coders. This can be easily verified by noting that applying a causal Wiener filter [119], which violates the PR constraint, to the reconstructed signal in a PR converter, is guaranteed to reduce distortion without introducing delay.

In the *subband coding* (SBC) literature, causal (zero-delay[8]) transform coders were first proposed in [120] by Habibi and Hershel, as an alternative to principal component transform coders (such as the *Karhunen-Loève Transform*, KLT, see [121]). Unlike KLT coders, causal transform coders use only triangular matrices for analysis and synthesis. The cost of achieving causality, in this case, is quantization noise amplification. The latter arises from the fact that triangular matrices cannot be unitary, and thus are not energy preserving (except for the identity matrix). Feeding the quantization error associated with one transform coefficient to the next coefficient *before* it is quantized, in a sequential fashion, reduces quantization noise gain, improving rate-distortion performance. This technique can be seen as a special case of feedback in the general architecture depicted in Fig. 1.4. Using the Linear Model (see Section1.2.2), it was shown by Phoong and Lin in [57] that careful design of the linear feedback component in a causal transform coder can, at high rates, bring the theoretical quantization noise gain down to unity. In such cases, the performance of causal transform coding equals that of KLT [57]. Notice that this is analogous

[8]The requirement of zero-delay is stronger than causality: a system can be causal and yet introduce arbitrary delay. Nevertheless, the term "causal transform coder" is commonly used to refer to zero-delay transform coders.

to the results mentioned above regarding biorthogonal filter banks as reported in [28]. The extension of causal transform coders to general subband coders is also discussed in [57]. The author is unaware of any other paper analyzing the design of SBCs. It is also important to note that the linear model analysis carried out in [57] and [28] assumes that fed-back quantization errors negligible variance. Thus, it is accurate only at high rates.

From a rate-distortion theoretical perspective, there exist partial results on the *optimal performance theoretically attainable* (OPTA) with zero-delay codes and causal source coding. Ericson [122], and Gaarder and Slepian [123, 124], have shown that, for i.i.d. sources, and under fixed-rate and zero-delay constraints, optimal rate-distortion performance is achieved by PDF-optimized scalar quantization. Other results have been obtained considering the less restrictive notion of *causality* instead of the requirement of zero end-to-end delay. This notion receives the name *causal source coding*, as introduced by Neuhoff and Gilbert in [125]. Under this concept, an encoder-decoder pair is deemed causal if the reconstruction of the current sample in the decoder is a function of *only* the current and past samples of the source. Notice that this definition allows for arbitrary delays in the entropy coding of the quantized samples. It is shown in [125] that for memoryless sources, the OPTA is achieved by time sharing between, at most, two entropy-constrained optimal scalar quantizers. It was later shown by Linder and Zamir that, for high rates, the cost of requiring causality in source-coding is approximately $0.254$ bis/sample with respect to Shannon's RDF [70]. It is also known that, for any source and at any rate-regime, the mutual information rate across an *additive white Gaussian noise* (AWGN) channel is not more than 0.5 bits/sample above the corresponding rate-distortion function [126]. By considering the use of subtractively-dithered scalar quantizers and entropy coding, this yields an upper bound for the OPTA in causal source coding at $0.754$ bits/sample above the non-causal RDF. However, it is unknown whether causal source coders can outperform this bound.

## 1.3 Overview of the Main Contributions

The main contributions of this thesis are as follows:

### 1.3.1 A Two-Parameter Frequency-Weighted MSE

The first contribution of this work is an extension of the MSE criterion to better address perceptual phenomena such as those shown in Fig. 1.2, and to account for *signal transfer function constraints* such as those discussed in Section 1.1.2. This measure consists of a weighted sum of the variance of the MSE component which is uncorrelated to the source, on the one hand, and the remainder of the MSE, on the

other. We give a formal definition to this distortion metric in the following.

For a random scalar source $\mathrm{x}$ reconstructed as $\mathrm{y}$ with reconstruction error $\mathrm{z} \triangleq \mathrm{y}-\mathrm{x}$, the mean squared error, $\sigma_{\mathrm{z}}^2 = \mathrm{E}\left[\mathrm{z}^2\right]$, can always be decomposed into two terms, namely, *source-uncorrelated error*

$$\mathrm{z} - \frac{\sigma_{\mathrm{z,x}}}{\sigma_{\mathrm{x}}^2}\,\mathrm{x}$$

and *source-parallel error*

$$\frac{\sigma_{\mathrm{z,x}}}{\sigma_{\mathrm{x}}^2}\,\mathrm{x}\,.$$

The variances of each of the above error components yields the *source-uncorrelated distortion*

$$D^{\perp} \triangleq \sigma_{\mathrm{z}}^2 - \frac{\sigma_{\mathrm{z,x}}^2}{\sigma_{\mathrm{x}}^2} \tag{1.4a}$$

and the *source-parallel distortion*

$$D^{\parallel} \triangleq \frac{\sigma_{\mathrm{z,x}}^2}{\sigma_{\mathrm{x}}^2}, \tag{1.4b}$$

such that

$$MSE = D^{\perp} + D^{\parallel} \tag{1.5}$$

As an extension of the MSE, the *Weighted Correlation Mean Squared Error* (WCMSE) between $\mathbf{x}$ and $\mathbf{y}$ is defined in this thesis as:

$$D_{a,b}(\mathbf{x},\mathbf{y}) \triangleq aD^{\perp} + bD^{\parallel}, \tag{1.6}$$

where $a$, $b$ are real positive coefficients. In the particular case of scalar random sources, $D^{\perp}$ and $D^{\parallel}$ are as in (1.4). For a w.s.s. random process source $\{\mathrm{x}(k)\}$ with reconstruction $\{\mathrm{y}(k)\} = \{\mathrm{x}(k) + \mathrm{z}(k)\}$,

$$D^{\perp} \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[S_{\mathrm{z}}(\mathrm{e}^{j\omega}) - \frac{\left|S_{\mathrm{z\,x}}(\mathrm{e}^{j\omega})\right|^2}{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}\right]d\omega, \tag{1.7}$$

$$D^{\parallel} \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\left|S_{\mathrm{z,x}}(\mathrm{e}^{j\omega})\right|^2}{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}d\omega. \tag{1.8}$$

where $S_{\mathrm{x}}$ and $S_{\mathrm{z}}$ are the *power spectral densities* (PSD) of $\{\mathrm{x}(k)\}$ and $\{\mathrm{z}(k)\}$, respectively, and $S_{\mathrm{z\,x}}(\mathrm{e}^{j\omega})$ is the cross-spectral density between $\{\mathrm{z}(k)\}$ and $\{\mathrm{x}(k)\}$. For a vector random source $\mathbf{x} \in \mathbb{R}^N$ having covariance matrix $\boldsymbol{K}_{\mathbf{x}}$, we have

$$D^{\perp} \triangleq \frac{1}{N}\,\mathrm{tr}\left\{\boldsymbol{K}_{\mathbf{z}}\right\} - \frac{1}{N}\,\mathrm{tr}\left\{\boldsymbol{K}_{\mathbf{z,x}}\boldsymbol{K}_{\mathbf{x}}^{-1}\boldsymbol{K}_{\mathbf{z,x}}^{T}\right\} \tag{1.9}$$

$$D^{\parallel} \triangleq \frac{1}{N}\,\mathrm{tr}\left\{\boldsymbol{K}_{\mathbf{z,x}}\boldsymbol{K}_{\mathbf{x}}^{-1}\boldsymbol{K}_{\mathbf{z,x}}^{T}\right\}, \tag{1.10}$$

where $\mathbf{z} = \mathbf{y} - \mathbf{x}$ and $\boldsymbol{K}_{\mathbf{z},\mathbf{x}}$ is the cross-covariance matrix between $\mathbf{z}$ and $\mathbf{x}$. Notice that setting $a = b = 1$ yields the standard MSE criterion.

Reconstruction errors produced by linear processing, such as filtering, are part of the source-parallel distortion term. These have no impact on the source-uncorrelated term. Thus, by choosing $b > a$, it is possible to assign a larger cost to deviations of the signal transfer function of an ED pair from a target transfer function. This allows WCMSE to take account of, for example, the perceptually greater impact of linear distortion in images such as those illustrated in Fig. 1.2, or signal transfer function constraints such as those described in Section 1.1.2. In particular, letting $b \to \infty$ yields a distortion metric which is in agreement with the situation where linear distortion is not tolerated. The design of source coder-decoder pairs that minimize the bit-rate for such a distortion metric will yield optimal unity transfer function (i.e., perfect reconstruction) source coders [9].

For the case of random processes and random vectors, it is straightforward to combine the WCMSE with frequency weighting. To be more precise, if the frequency sensitivity to each distortion term is the same, say $P(\mathrm{e}^{j\omega})$, then the frequency-weighted WCMSE becomes

$$a\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|P(\mathrm{e}^{j\omega})\right|^2 \left[S_{\mathrm{z}}(\mathrm{e}^{j\omega}) - \frac{S_{\mathrm{x,z}}(\mathrm{e}^{j\omega})}{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}\right] d\omega + b\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|P(\mathrm{e}^{j\omega})\right|^2 \frac{S_{\mathrm{x,z}}(\mathrm{e}^{j\omega})}{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} d\omega \tag{1.11}$$

Obviously, frequency weighted WCMSE includes frequency weighted MSE as a special case. Furthermore, provided appropriate values for $a$ and $b$ are chosen, frequency weighted WCMSE will be superior to frequency weighted MSE in all applications where source-correlated distortion has a different impact than linear distortion.

### 1.3.2 WCMSE Optimal Frequency Responses for Scalar Feedback Quantizers

The second contribution of this work is the derivation of the frequency response of the filters in a general feedback scalar quantization system that minimize the frequency weighted WCMSE, for a given quantizer SNR constraint, and for any choice of weights $a$, $b$. This optimization problem is solved for various combinations of filters being fixed and given as listed in Table 1.1. Since the WCMSE is novel to this thesis, the solutions are new and include MSE-optimal filters as special cases. For example, for the last problem in Table 1.1, the results presented here include *the first optimization-based* derivation of the optimal filters characterized by Zamir, Kochman and Erez in [15].

[9] With the choice $a = 0$, any optimal ED pair would yield no linear distortion, but the source-uncorrelated distortion would be unbounded, which makes this extreme case of little practical interest.

These results are then applied to design a family of scalar feedback quantizers whose WCMSE decays exponentially with the oversampling ratio, for a fixed quantizer SNR, assuming overload errors are negligible. If a subtractively dithered quantizer is utilized, then the noise model is exact, and the SNR constraint can be directly related to the bit-rate if entropy coding is used, regardless of the number of quantization levels. It is shown that in optimal feedback quantizers with entropy coded dithered quantization, the WCMSE decays with the oversampling ratio as $2^{-1.75\lambda}$. In the case of fixed-rate quantization, the SNR is related to the number of quantization levels, and hence to the bit-rate, when overload errors are negligible. It is shown that, for sources with unbounded support, the latter condition is violated for oversampling ratios sufficiently large. By deriving an upper bound on the contribution of overload errors to the total WCMSE, a lower bound for the decay rate of the WCMSE as a function of the oversampling ratio is found for fixed-rate quantization. To the best of the author's knowledge, this is the first bound of this type that takes into account overload errors. This makes the result applicable to the characterization of the oversampling efficiency of feedback quantizers at encoding signals having unbounded support.

### 1.3.3 The Rate-Distortion Function for Gaussian Sources with WCMSE as Distortion Measure

The third contribution of this thesis is the derivation of the rate-distortion function for Gaussian sources, when WCMSE is used as the distortion metric. This RDF, denoted by $R_{a,b}(D)$, yields the well known water filling equations when $a = b = 1$, and the *RDF for source uncorrelated distortions* $R^{\perp}(D)$, which was recently introduced by the author in [127], when $a = 1$ and $b \to \infty$.[10] In addition, $R^{\perp}(D)$ is characterized for the case in which there exists LTI feedback between the output and the input of the ED pair.

### 1.3.4 Using Realizations of the RDF to Design Optimal Source Coders

It is shown in this thesis that, under the Linear Model, the WCMSE-optimal filters in feedback quantizers *having three degrees of freedom*, subject to a quantizer SNR constraint, are also WCMSE-optimal when the constraint is the *end-to-end mutual information rate* and the quantizer is substituted by an AWGN channel. These results provide conditions under which the knowledge of a realization of the underlying

[10] It may seem at first surprising that fixing $a = 1$ and letting $b \to \infty$ yields only source-uncorrelated distortion. This apparent contradiction is clarified by noting that a large value of the weight $b$ implies that, in order to minimize $D_{a,b}$ for a given rate, source-parallel distortion should be small, since it is more expensive than source-uncorrelated distortion. In the limit as $b \to \infty$, source-parallel distortion is infinitely expensive, and thus the minimization of $D_{a,b}$ can only allow for source-uncorrelated distortion.

rate-distortion function can be used as a guideline to design optimal ED pairs. Necessary and sufficient conditions for the equivalence between the quantizer-SNR-constrained optimization problem and the Mutual-Information-Constrained optimization problem are found. This insight is then further extended to other implementations (filter banks and transform coders with feedback) of the general architecture shown in Fig. 1.4.

### 1.3.5 Results on the Causal Quadratic Gaussian Rate-Distortion Function

By using $R^{\perp}(D)$ as a starting point, an iterative method is found for obtaining an upper bound on the information-theoretic causal RDF for Gaussian stationary sources under the MSE distortion criterion. It is shown that this method always converges, and that the bound, thus obtained, is tighter than $0.5$ bits/sample. Moreover, if there exists a realization of the information-theoretic causal RDF in which the reconstruction error is jointly stationary with the source, then this bound coincides with the information-theoretic causal RDF. In addition, the method yields the frequency response of the filters in a *causal* scalar feedback quantizer which achieves a rate $0.254$ bits/sample above the latter bound. This constitutes an upper bound on the optimal performance theoretically attainable by any causal source coder for stationary Gaussian sources under the MSE distortion criterion.

### 1.3.6 Summary of the Main Contributions

Summarizing, the main contribution of this thesis are:

1. A novel extension of the MSE beyond frequency weighting, named WCMSE, is presented, to incorporate signal transfer function constraints.

2. In Chapter 3, feedback scalar quantization optimization problems with architectural constraints are solved, using the Linear Model and the WCMSE as distortion metric and the quantizer SNR as the bit-rate constraint. These results are applied to design a family of scalar feedback quantizers whose WCMSE decays exponentially with the oversampling ratio, for a fixed quantizer SNR.

3. In Chapter 4, the RDF for the WCMSE as the distortion metric is introduced and then completely characterized for Gaussian sources. This RDF is then extended to cases where there exists LTI feedback between reconstructed signal and source. The implications of this result for networked control theory are also discussed.

4. In Chapter 5, conditions and principles are found upon which end-to-end-mutual information and quantizer SNR are equivalent constraints in the design of encoder-decoder pairs for minimum

WCMSE. It is also shown how this result can help in the design of optimal subband coders.

5. In Chapter 6, it is shown that, for stationary Gaussian sources with memory, an upper bound can be obtained for the *causal* WCMSE rate-distortion function by means of an iterative procedure. This bound is equal to the causal RDF if the latter admits a realization in which the reconstruction error is jointly stationary with the source.

## 1.4 Associated Publications

Most of the results in this thesis have not been published. However, some of these results are based on earlier work that has been published as indicated in the following:

### Journal Publications

M. S. Derpich, E. Silva, D. E. Quevedo, and G. C. Goodwin, "On optimal perfect reconstruction feedback quantizers," *IEEE Trans. Signal Process*., vol. 56, no. 8, Part 2, pp. 3871–3890, August 2008.

### Conference Publications

1. M. S. Derpich, D. E. Quevedo, and G. C. Goodwin, "Probability of interpolation for a mute sample interpolative A/D converter with horizon-length two," in *Proc. IEEE TENCON*, November 2005, pp. 1–6.

2. M. S. Derpich, D. E. Quevedo, G. C. Goodwin, and A. Feuer, "Quantization and sampling of not necessarily band-limited signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*., vol. 3, Toulouse, France, May 2006, pp. 396–399.

3. M. S. Derpich, D. E. Quevedo, and G. C. Goodwin, "Optimal AD-conversion via sampled-data receding horizon control theory." in *Proc. Chinese Control Conf.*, Harbin, August 2006, pp. 1417–1422

4. G. C. Goodwin, M. S. Derpich, and D. E. Quevedo, "Efficient data representations for signal processing and control: "making most of a little"," in *Proc. Chinese Control Conf.*, August 2006, pp. PL-17–PL-37.

5. E. I. Silva, G. C. Goodwin, D. E. Quevedo, and M. S. Derpich, "Optimal noise shaping for networked control systems," in *Proc. Europ. Contr. Conf.*, Kos, Greece, July 2007.

6. M. S. Derpich, D. E. Quevedo, and G. C. Goodwin, "Conditions for optimality of scalar feedback quantization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Las Vegas, NV, USA, 2008, pp. 3749–3752.

7. M. S. Derpich, J. Østergaard, and G. C. Goodwin, "The quadratic Gaussian rate- distortion function for source uncorrelated distortions," in *Proc. Data Compression Conf.*, Snowbird, UT, USA, March 2008, pp. 73–82.

8. E. I. Silva, M. S. Derpich, J. Østergaard, and D. E. Quevedo, "Simple coding for achieving mean square stability over bit-rate limited channels," in *Proc. IEEE Conf. Decis. Contr.*, Cancún, Mexico, December 2008, pp. 3871–3890.

## Technical Reports

1. M. S. Derpich, G. C. Goodwin, and D. E. Quevedo, "Efficient analog-to-digital conversion via interpolation and quantized moving horizon control," 2005.

2. M. S. Derpich, E. I. Silva, D. E. Quevedo, and G. C. Goodwin, "Optimal noise- shaping DPCM," available from http://msderpich.no-ip.org.

3. M. Derpich, J. Østergaard, and D. Quevedo, "Achieving the quadratic Gaussian rate-distortion function for source uncorrelated distortions," (available at http://arxiv.org/abs/0801.1718v3 ).

# Chapter 2

# Preliminaries

*A good notation has a subtlety and suggestiveness which at times make it seem almost like a live teach.*

*Attributed to Bertrand Russell, British mathematician and philosopher.*

*My greatest concern was what to call it. I thought of calling it "information", but the word was overly used, so I decided to call it "uncertainty". When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage".*

*Claude Shannon, United States electronic engineer and mathematician.*

## 2.1 Notation

- $\mathbb{N}$ is the set of natural numbers.
- $\mathbb{Z}$ is the set of integer numbers.
- $\mathbb{R}$ is the set of real numbers.
- $\mathbb{C}$ is the set of complex numbers.
- $x$, $X$, lower and upper case italic letters are used for scalars.
- $\boldsymbol{x}$ lower case italic bold letters are used for vectors.
- $\boldsymbol{X}$ uppercase italic bold letters are used for matrices.
- $\{x(k)\}$ is used for infinite length sequences.

- $x^k$ is a short-hand notation for the semi-infinite length sequence $\{x(i)\}_{i=-\infty}^{k}$, $k \in \mathbb{Z}$.
- $x_j^k$ is a short-hand notation for the finite length sequence $\{x(i)\}_{i=j}^{k}$, $j, k \in \mathbb{Z}$.
- $x^*$ denotes the complex conjugate of $x$.
- $\boldsymbol{X}^T$ denotes the transpose of the matrix $\boldsymbol{X}$.
- $\boldsymbol{X}^H$ denotes the Hermitian (conjugate-transpose) of the matrix $\boldsymbol{X}$, i.e., $\boldsymbol{X}^H = (\boldsymbol{X}^T)^*$.
- $\boldsymbol{X}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\boldsymbol{X}$.
- $\mathrm{tr}\,\{\boldsymbol{X}\}$ denotes the trace of a matrix $\boldsymbol{X}$.
- $|\boldsymbol{X}|$ denotes the determinant of the matrix $\boldsymbol{X}$.
- $|\boldsymbol{X}|_{HS}$ denotes the *weak matrix norm* of $\boldsymbol{X}$, see Definition 2.5 below.
- $\|\boldsymbol{X}\|$ denotes the *strong matrix norm* of $\boldsymbol{X}$, see Definition 2.4 below.
- $\boldsymbol{X}_\ell \sim \boldsymbol{Y}_{\boldsymbol{y}}$ denotes asymptotic equivalence between the sequences of matrices $\boldsymbol{X}_\ell$ and $\boldsymbol{Y}_\ell$, see Definition 2.7 below.
- $\lambda_i(\boldsymbol{X})$ denotes the $i$-th eigenvalue of the matrix $\boldsymbol{X}$, where $\lambda_i(X) \geq \lambda_j\{\boldsymbol{X}\}$ if $i > j$.
- $\mathrm{diag}\{x_k\}$ is a diagonal square matrix (of appropriate dimension), with diagonal elements $x_k$.
- $\boldsymbol{I}$ is the identity matrix (of appropriate dimension).
- $\mathrm{x}, \mathbf{x}, \mathbf{X}$ non-italic fonts for random scalars, vectors and matrices.
- $\mathrm{E}\,[\cdot]$ denotes the expectation operator.
- $\sigma_{\mathrm{x}}^2 = \mathrm{E}\,[\mathrm{x}\,\mathrm{x}^*]$ $(= \mathrm{E}\,[\mathrm{x}(k)\,\mathrm{x}(k)^*])$ is the variance of the zero-mean random variable x (or the zero-mean w.s.s. process $\{\mathrm{x}(k)\}$).
- $\sigma_{\mathrm{x,y}} = \mathrm{E}\,[\mathrm{x}\,\mathrm{y}^*]$ is the covariance between the two zero-mean scalar random variables x and y.
- $\boldsymbol{K}_{\mathbf{x}} = \mathrm{E}\left[\mathbf{x}\mathbf{x}^H\right]$ is the covariance matrix of the random vector $\mathbf{x}$.
- $\boldsymbol{K}_{\mathbf{z},\mathbf{x}} = \mathrm{E}\left[\mathbf{z}\mathbf{x}^H\right]$ denotes the cross-covariance matrix between the random vectors $\mathbf{z}$ and $\mathbf{x}$.
- $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$ denotes the *power spectral density* (PSD) of the wide sense stationary (w.s.s.) random process $\{\mathrm{x}(k)\}$.

- $\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega}) = \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}$ is the square root of the PSD of the w.s.s. random process $\{\mathrm{x}(k)\}$.
- $L^2$ and $L^1$ are sets of all complex-valued functions that are square integrable and absolutely integrable over $[-\pi, \pi]$, respectively.
- $\ell^2$ and $\ell^1$ are sets of all complex-valued square integrable and absolutely integrable sequences, respectively.
- $\langle \cdot, \cdot \rangle$ denotes the inner product between its arguments In particular,
  - If $F(\mathrm{e}^{j\omega})$, $G(\mathrm{e}^{j\omega}) \in L^2$, then $\langle F, G \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(\mathrm{e}^{j\omega}) G(\mathrm{e}^{j\omega})^* d\omega$.
  - If $f(\cdot), g(\cdot) \in L^2$, then $\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) g^*(x) dx$.
- $\|\cdot\|$ denotes the 2-norm $\sqrt{\langle \cdot, \cdot \rangle}$.
- $\|\cdot\|_\infty$ denotes the $\infty$-norm. For a sequence $\{x(k)\}$, $\|\cdot\|_\infty$ corresponds to the $\ell_\infty$ norm $\|x\|_\infty = \max_{k \in \mathbb{Z}} \{|x(k)|\}$. For a function of a continuous variable $f : [a, b] \to \mathbb{R}$, it corresponds to the $L_\infty$ norm $\|f\|_\infty = \max_{x \in [a,b]} \{f(x)\}$.
- $\mathcal{N}_f$ denotes the null-space of the function, mapping or transformation $f$, i.e., the set of arguments whose image though $f$ is zero. (With some abuse of notation, for a discrete-time Fourier transform $F(\mathrm{e}^{j\omega})$ we write $\mathcal{N}_F \triangleq \{\omega \in [-\pi, \pi] : F(\mathrm{e}^{j\omega}) = 0\}$.)

## 2.2 Definitions

To simplify notation, we introduce the operator $(\cdot)^{\sim 1}$, defined as follows:

$$F(\mathrm{e}^{j\omega})^{\sim 1} = \begin{cases} F(\mathrm{e}^{j\omega})^{-1} & , \quad \forall \omega \notin \mathcal{N}_F \\ \natural & , \quad \forall \omega \in \mathcal{N}_F, \end{cases} \tag{2.1}$$

where $F : \mathbb{C} \to \mathbb{R}$ is any given function and $\natural$ denotes any arbitrary and positive bounded value.

**Definition 2.1** (Gateaux Differential [128])**.** *Let $\mathcal{X}$ be a vector space and $\mathscr{V}$ a functional from $\mathcal{X}$ to $\mathbb{R}$. The* Gateaux differential *of $\mathscr{V}$ at $f \in \mathcal{X}$ with increment $h \in \mathcal{X}$ is defined as*

$$\delta \mathscr{V}(f; h) \triangleq \frac{d}{d\alpha} \mathscr{V}(f + \alpha h)\Big|_{\alpha = 0}, \tag{2.2}$$

*if the above derivative exists.*

**Definition 2.2** (Similarly/Oppositely Functionally Related)**.** We say that two functions $\phi, \psi : [a, b] \to \mathbb{R}$ are *similarly functionally related* iff there exists a monotonically increasing function $G(\cdot)$ such that $\phi(x) = G(\psi(x))$, for all $x \in [a, b]$, and write $\phi \uparrow\uparrow \psi$. Similarly, if there exists a monotonically decreasing function $G(\cdot)$ such that $\phi(x) = G(\psi(x))$, for all $x \in [a, b]$, we say that $\phi$ and $\psi$ are *oppositely functionally related*, and write $\phi \uparrow\downarrow \psi$. ▲

**Definition 2.3** (Almost Constant Function)**.** A function $f : [-\pi, \pi] \to \mathbb{R}$ is said to be almost constant iff

$$\int_{-\pi}^{\pi} \left| f(x) - \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\omega) d\omega \right| dx = 0. \tag{2.3}$$

▲

**Definition 2.4** (Strong Matrix Norm [129])**.** *The* strong norm *of a matrix $\boldsymbol{A}$, denoted by $\|\boldsymbol{A}\|$, is defined as*

$$\|\boldsymbol{A}\| \triangleq \max_{\boldsymbol{z}:\boldsymbol{z}^H\boldsymbol{z}=1} \left[ \boldsymbol{z}^H \boldsymbol{A}^H \boldsymbol{A} \boldsymbol{z} \right]^{1/2} = \max_i |\lambda_i(\boldsymbol{A})|^2 \tag{2.4}$$

▲

**Definition 2.5** (Weak Matrix Norm [129])**.** *The* Hilbert-Schmidt *or* weak norm *of a matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, denoted by $|\boldsymbol{A}|_{HS}$, is defined as*

$$|\boldsymbol{A}|_{HS} \triangleq \left( \frac{1}{N} \operatorname{tr} \left\{ \boldsymbol{A}^H \boldsymbol{A} \right\} \right)^{1/2} = \left( \frac{1}{N} \sum_{i=1}^{N} |\lambda_i(\boldsymbol{A})|^2 \right)^{1/2}. \tag{2.5}$$

▲

**Definition 2.6** (Wiener Class Toeplitz and Circulant Matrices [129])**.** *For a given function $f \in L^1 : [-\pi, \pi] \to \mathbb{C}$, having absolutely summable inverse discrete-time Fourier Transform, the Toeplitz matrix $\boldsymbol{T}_\ell(f) \in \mathbb{R}^{\ell \times \ell}$ is defined element-wise as*

$$[\boldsymbol{T}_\ell(f)]_{m,n} \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\omega)\, \mathrm{e}^{-j[n-m]\omega} d\omega, \quad m, n = 0, 1, \ldots, \ell - 1. \tag{2.6}$$

*Similarly, the circulant matrix $\boldsymbol{C}_\ell(f) \in \mathbb{R}^{\ell \times \ell}$ is defined as the circulant matrix whose top row is given by*

$$[\boldsymbol{C}_\ell(f)]_{1,m+1} \triangleq \frac{1}{\ell} \sum_{k=0}^{\ell-1} f(2\pi k/\ell)\, \mathrm{e}^{j2km\pi/\ell}, \quad m = 0, 1, \ldots, \ell - 1. \tag{2.7}$$

▲

**Definition 2.7** (Asymptotically Equivalent Sequences of Matrices [129])**.** *Two sequences of $\ell \times \ell$ matrices $\{\boldsymbol{A}_\ell\}_{\ell=1}^{\infty}$, $\{\boldsymbol{B}_\ell\}_{\ell=1}^{\infty}$ are said to be* asymptotically equivalent*, denoted by*

$$\boldsymbol{A}_\ell \sim \boldsymbol{B}_\ell,$$

*if*

1. *$\boldsymbol{A}_\ell$ and $\boldsymbol{B}_\ell$ are uniformly bounded in the strong (and hence in weak) norm, i.e.,*

$$\|\boldsymbol{A}_\ell\|, \|\boldsymbol{B}_\ell\| \leq M < \infty, \ell = 1, 2, \ldots, \tag{2.8}$$

   *and*

2. *$\boldsymbol{D}_\ell \triangleq \boldsymbol{A}_\ell - \boldsymbol{B}_\ell$ goes to zero in the weak norm as $\ell \to \infty$, i.e.,*

$$\lim_{\ell \to \infty} |\boldsymbol{A}_\ell - \boldsymbol{B}_\ell|_{HS} = \lim_{\ell \to \infty} |\boldsymbol{D}|_{HS} = 0 \tag{2.9}$$

▲

## 2.3 Basic Information Theoretical Concepts and Results

The following is a brief list of some of the information theoretical quantities and properties that will be useful in the derivations carried out in this thesis. For proofs and insightful descriptions about these concepts and results, the reader is referred to, e.g., [7, 9, 63].

### 2.3.1 Entropy

**Definition 2.8** (Entropy of a Discrete Random Variable)**.** *The entropy of a discrete random variable* $\mathrm{x}$ *taking values from a countable set $\mathbb{X}$, with probability mass function $p_{\mathrm{x}}(\cdot)$, is defined as*

$$H(\mathrm{x}) \triangleq -\sum_{x \in \mathbb{X}} p_{\mathrm{x}}(x) \log(p_{\mathrm{x}}(x)) \tag{2.10}$$

▲

If the $\log(\cdot)$ in (2.10) is taken to be $\log_2(\cdot)$, then the units of the entropy of $\mathrm{x}$ are **bits**. If instead we use $\ln(\cdot)$ in place of $\log(\cdot)$ in (2.10), then the units of $H(\mathrm{x})$ are **nats**. For any discrete random variable $\mathrm{x}$,

$$H(\mathrm{x}) \geq 0. \tag{2.11}$$

**Definition 2.9** (Entropy of an Ensemble of Discrete Random Variables)**.** *The entropy of an ensemble of discrete random variables* $\{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N\}$*, each of them taking values from countable sets* $\{\mathbb{X}_i\}_{i=1}^N$*, with joint probability mass function* $p_{\mathrm{x}_1,\ldots,\mathrm{x}_N}(\cdot)$*, is defined as*

$$H(\mathrm{x}) \triangleq -\sum_{x_1\in\mathbb{X}_1}\cdots\sum_{x_N\in\mathbb{X}_N} p_{\mathrm{x}_1,\ldots,\mathrm{x}_N}(x_1,\ldots,x_N)\log(p_{\mathrm{x}_1,\ldots,\mathrm{x}_N}(x_1,\ldots,x_N)) \tag{2.12}$$

▲

**Definition 2.10** (Conditional Entropy for Discrete Random Variables)**.** *If* $(\mathrm{x},\mathrm{y}) \sim p_{\mathrm{x},\mathrm{y}}(\cdot,\cdot)$*, the* ***conditional entropy*** $H(\mathrm{x}\,|\,\mathrm{y})$ *is defined as*

$$H(\mathrm{y}\,|\,\mathrm{x}) \triangleq \sum_{x\in\mathbb{X}} p_{\mathrm{x}}(x) H(\mathrm{y}\,|\,\mathrm{x}=x) \tag{2.13}$$

$$= \sum_{x\in\mathbb{X}} p_{\mathrm{x}}(x) \sum_{y\in\mathbb{Y}} p_{\mathrm{y}\,|\,\mathrm{x}}(y|x)\log(p_{\mathrm{y}\,|\,\mathrm{x}}(y|x)) \tag{2.14}$$

The notion of entropy can also be applied to continuous random variables:

**Definition 2.11.** *The* ***differential entropy*** *of a continuous random variable* x *with PDF* $f_{\mathrm{x}}(\cdot)$ *is defined as*

$$h(\mathrm{x}) \triangleq -\int_{x\in\mathbb{X}} f_{\mathrm{x}}(x)\log(f_{\mathrm{x}}(x))dx, \tag{2.15}$$

*where* $\mathbb{X}$ *is the support of* $f_{\mathrm{x}}(\cdot)$*.*

Unlike discrete entropy, differential entropy can be negative. Indeed, $h(\mathrm{x})$ is differential in the sense that it is relative to the entropy of a random variable u, distributed uniformly over a unit-length interval. Such a random variable will have zero differential entropy. Thus, if the differential entropy x is 2 bits, this means that its entropy is 2 bits higher than that of the uniformly distributed random variable u.

It is easy to verify that the differential entropy of a Gaussian scalar random variable $\mathrm{x}_G \sim \mathcal{N}(0, \sigma^2_{\mathrm{x}_G})$ is

$$h(\mathrm{x}_G) = \frac{1}{2}\log_2(2\pi\,\mathrm{e}\,\sigma^2_{\mathrm{x}_G}) \quad \text{bits.} \tag{2.16}$$

**Definition 2.12.** *The* ***joint*** *differential entropy of a continuous random vector* $\mathbf{x}$ *whose elements have joint PDF* $f_{\mathbf{x}}(\cdot)$ *is defined as*

$$h(\mathbf{x}) \triangleq -\int_{\boldsymbol{x}\in\mathbb{X}} f_{\mathbf{x}}(\boldsymbol{x})\log(f_{\mathbf{x}}(\boldsymbol{x}))d\boldsymbol{x}, \tag{2.17}$$

*where* $\mathbb{X}$ *is the support of* $f_{\mathbf{x}}(\cdot)$*.*

For two continuous random vectors $\mathbf{x}, \mathbf{y}$ having joint PDF $f_{\mathbf{x},\mathbf{y}}(\cdot,\cdot)$, the **conditional differential entropy** $h(\mathbf{y}|\mathbf{x})$ is defined as

$$h(\mathbf{y}|\mathbf{x}) \triangleq -\int f_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}) \log f_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} d\boldsymbol{x} \tag{2.18}$$

It is often useful to decompose $h(\mathrm{y}\,|\,\mathrm{x})$ as

$$h(\mathbf{y}|\mathbf{x}) = h(\mathbf{x},\mathbf{y}) - h(\mathbf{x}) \tag{2.19}$$

which holds if and only if each of the differential entropies in (2.19) is bounded.

**Property 2.1.** $h(\mathbf{y}) \geq h(\mathbf{y}|\mathbf{x})$*, with equality if and only if* $\mathbf{x}$ *and* $\mathbf{y}$ *are independent.*

**Property 2.2.** $h(\mathbf{y}|\mathbf{x}) \geq h(\mathbf{y}|\mathbf{x},\mathbf{z})$*, with equality if and only if* $\mathbf{y}$ *and* $\mathbf{z}$ *are conditionally independent given* $\mathbf{x}$*. (See also Definition 2.18 on page 39.)*

**Property 2.3.** $h(\mathbf{y},\mathbf{x}) \leq h(\mathbf{x}) + h(\mathbf{y})$*, with equality if and only if* $\mathbf{x}$ *and* $\mathbf{y}$ *are independent.*

**Property 2.4.** *If* $\mathrm{x}$ *is a scalar random variable and* $c$ *is some arbitrarily constant,*

$$h(\mathrm{x}+c) = h(\mathrm{x}). \tag{2.20}$$

**Property 2.5.** *If* $\mathrm{x}$ *is a scalar random variable and* $c \neq 0$ *is some arbitrarily constant, then*

$$h(c\,\mathrm{x}) = h(\mathrm{x}) + \log|c| \tag{2.21}$$

**Property 2.6.** *If* $f(\cdot)$ *is any given deterministic function, then*

$$h(\mathbf{x} + f(\mathbf{y})|\mathbf{y}) = h(\mathbf{x}|\mathbf{y}). \tag{2.22}$$

**Property 2.7.** *If* $\mathbf{x} \in \mathbb{R}^N$ *is random vector and* $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ *is some arbitrarily matrix, then*

$$h(\boldsymbol{M}\mathbf{x}) = h(\mathbf{x}) + \log|\det(\boldsymbol{M})| \tag{2.23}$$

**Property 2.8.** *(Chain rule for differential entropy)*

$$h(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_N) = \sum_{k=1}^{N} h(\mathrm{x}_k\,|\,\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_{k-1}), \tag{2.24}$$

$$h(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_N\,|\,\mathrm{z}) = \sum_{k=1}^{N} h(\mathrm{x}_k\,|\,\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_{k-1},\mathrm{z}). \tag{2.25}$$

The **differential entropy per dimension** of a random vector $\mathbf{x} \in \mathbb{R}^N$ is denoted by

$$\bar{h}(\mathbf{x}) = \frac{1}{N} h(\mathbf{x}) \tag{2.26}$$

**Fact 2.1** (From [7, Theorem 14])**.** *If an ensemble of random variables having entropy $h_1$ per degree of freedom is passed through a filter with frequency response $F(\mathrm{e}^{j\omega})$, then the output ensemble has differential entropy*

$$h_2 = h_1 + \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|F(\mathrm{e}^{j\omega})\right|^2 d\omega. \tag{2.27}$$

**Definition 2.13.** *The differential **entropy rate** of a random process $\{\mathrm{x}(k)\}_{k=1}^{\infty}$ is defined by*

$$\bar{h}(\{\mathrm{x}(k)\}) \triangleq \lim_{\ell\to\infty} \frac{1}{\ell} h(\mathrm{x}(1), \mathrm{x}(2), \ldots, \mathrm{x}(\ell)) = \lim_{\ell\to\infty} \frac{1}{\ell} h(\mathrm{x}_1^{\ell}), \tag{2.28}$$

*whenever the limit exists.*

**Fact 2.2.** *If $\{\mathrm{x}(k)\}_{k=-\infty}^{\infty}$ is a stationary process, then*

$$\bar{h}(\{\mathrm{x}(k)\}) = h(\mathrm{x}(k)|\ldots, \mathrm{x}(k-2), \mathrm{x}(k-1)) = h(\mathrm{x}(k)|\,\mathrm{x}^{k-1}) \tag{2.29}$$

**Fact 2.3.** *If $\{\mathrm{x}(k)\}$ is a Gaussian stationary process with PSD $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$, then*

$$\bar{h}(\{\mathrm{x}(k)\}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log\left(2\pi\,\mathrm{e}\, S_{\mathrm{x}}(\mathrm{e}^{j\omega})\right) d\omega \tag{2.30}$$

**Definition 2.14.** *The **relative entropy (or Kullback-Leibler distance)** $D(f\|g)$ between two PDFs $f$ and $g$ is defined by*

$$D(f\|g) \triangleq \int f \log \frac{f}{g} \tag{2.31}$$

### 2.3.2 Mutual Information

**Definition 2.15.** *The* mutual information $I(\mathrm{x};\mathrm{y})$ *between two random variables with joint PDF $f_{\mathrm{x},\mathrm{y}}(\cdot,\cdot)$ is defined as*

$$I(\mathrm{x};\mathrm{y}) \triangleq \int f_{\mathrm{x},\mathrm{y}}(x,y) \log \frac{f_{\mathrm{x},\mathrm{y}}(x,y)}{f_{\mathrm{x}}(x) f_{\mathrm{y}}(y)} dxdy \tag{2.32}$$

The mutual information $I(\mathrm{x};\mathrm{y})$ is a measure of the amount of information that any of the random variables involved contains about the other. From Definition (2.15) it is clear that

$$I(\mathrm{x};\mathrm{y}) = h(\mathrm{x}) - h(\mathrm{x}\,|\,\mathrm{y}) \tag{2.33}$$

$$= h(\mathrm{y}) - h(\mathrm{y}\,|\,\mathrm{x}) \tag{2.34}$$

$$= h(\mathrm{x}) + h(\mathrm{y}) - h(\mathrm{x},\mathrm{y}), \tag{2.35}$$

and that

$$I(\mathrm{x};\mathrm{y}) = I(\mathrm{y};\mathrm{x}). \tag{2.36}$$

Other well known properties are the following:

**Property 2.9.** *$D(f\|g) \geq 0$, with equality if and only if $f = g$ almost everywhere.*

**Property 2.10.** *$I(\mathrm{x};\mathrm{y}) \geq 0$, with equality if and only if $\mathrm{x}$ and $\mathrm{y}$ are independent.*

**Property 2.11.** *For random variables $\mathrm{x}$, $\mathrm{y}$ and $\mathrm{z}$,*

$$I(\mathrm{x};\mathrm{y}\,|\,\mathrm{z}) = h(\mathrm{x}\,|\,\mathrm{z}) - h(\mathrm{x}\,|\,\mathrm{y},\mathrm{z}) \tag{2.37}$$

*provided the differential entropies on the right-hand side of* (2.37) *are bounded.*

**Definition 2.16.** *The **mutual information per dimension** between two random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ is defined as*

$$\bar{I}(\mathbf{x};\mathbf{y}) \triangleq \frac{1}{N} I(\mathbf{x};\mathbf{y}). \tag{2.38}$$

**Definition 2.17.** *The **mutual information rate** between two jointly stationary random processes $\{\mathrm{x}(k)\}_{k=1}^{\infty}$ and $\{\mathrm{y}(k)\}_{k=1}^{\infty}$ is defined as*

$$\bar{I}(\{\mathrm{x}(k)\}\,;\{\mathrm{y}(k)\}) \triangleq \lim_{\ell\to\infty} \frac{1}{\ell} I(\mathrm{x}_1^{\ell};\mathrm{y}_1^{\ell}), \tag{2.39}$$

*provided the limit exists.* ▲

**Fact 2.4.** *If $\{\mathrm{x}(k)\}$ and $\{\mathrm{z}(k)\}$ are independent Gaussian stationary processes then*

$$\bar{I}(\{\mathrm{x}(k)\}\,;\{\mathrm{x}(k)+\mathrm{z}(k)\}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log\left(\frac{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + S_{\mathrm{z}}(\mathrm{e}^{j\omega})}{S_{\mathrm{z}}(\mathrm{e}^{j\omega})}\right) d\omega \tag{2.40}$$

**Definition 2.18.** *Random variables $\mathrm{x}$, $\mathrm{y}$, $\mathrm{z}$ are said to form a **Markov chain** in that order (denoted by $\mathrm{x} \to \mathrm{y} \to \mathrm{z}$) if the conditional distribution of $\mathrm{z}$ depends only on $\mathrm{y}$ and is conditionally independent of $\mathrm{x}$. Specifically, $\mathrm{x}$, $\mathrm{y}$, $\mathrm{z}$ form a Markov chain $\mathrm{x} \to \mathrm{y} \to \mathrm{z}$ if their joint PDF can be written as*

$$f_{\mathrm{x},\mathrm{y},\mathrm{z}}(x,y,z) = f_{\mathrm{x}}(x) f_{\mathrm{y}\,|\,\mathrm{x}}(y|x) f_{\mathrm{z}\,|\,\mathrm{y}}(z|y), \tag{2.41}$$

*or, equivalently, if*

$$f_{\mathrm{z}\,|\,\mathrm{x},\mathrm{y}}(z|x,y) = f_{\mathrm{z}\,|\,\mathrm{y}}(z|y) \tag{2.42}$$

**Property 2.12.** *Random variables* $\mathrm{x}$, $\mathrm{y}$ , $\mathrm{z}$ *form the Markov chain* $\mathrm{x} \to \mathrm{y} \to \mathrm{z}$ *if and only if* $\mathrm{z}$ *and* $\mathrm{x}$ *are conditionally independent given* $\mathrm{y}$*, i.e., if and only if*

$$f_{\mathrm{x,z\,|\,y}}(x, z|y) = f_{\mathrm{x\,|\,y}}(x|y) f_{\mathrm{z\,|\,y}}(z|y). \tag{2.43}$$

**Property 2.13.** *The Markov chain* $\mathrm{x} \to \mathrm{y} \to \mathrm{z}$ *implies* $\mathrm{z} \to \mathrm{y} \to \mathrm{x}$*.*

**Property 2.14.** *If* $\mathrm{z} = f(\mathrm{x})$*, where* $f(\cdot)$ *is some deterministic function, then* $\mathrm{x} \to \mathrm{y} \to \mathrm{z}$*.*

**Fact 2.5** (Data Processing Inequality)**.** *If* $\mathrm{x} \to \mathrm{y} \to \mathrm{z}$*, then* $I(\mathrm{x};\mathrm{y}) \geq I(\mathrm{x};\mathrm{z})$ *and* $I(\mathrm{y};\mathrm{z}) \geq I(\mathrm{x};\mathrm{z})$*.*

## 2.4 Scalar Memoryless Quantization

### 2.4.1 Uniform Scalar Quantization

A uniform scalar quantizer $\mathcal{Q}$ having $L$ levels and quantization interval $\Delta$ is defined by the following mapping:

$$\mathcal{Q}(v) = \arg\min_{\mu \in \mathbb{U}} |\mu - v|\,, \tag{2.44}$$

where the *quantization alphabet* $\mathbb{U}$ is given by

$$\mathbb{U} \triangleq \left\{ u_k = -\tfrac{L\Delta}{2} - \tfrac{\Delta}{2} + k\Delta,\ k = 1, 2, \ldots, L \right\}. \tag{2.45}$$

If $L$ is odd, then $0 \in \mathbb{U}$, and then $\mathcal{Q}$ is a *mid-step* quantizer. Else, if $L$ is even, $\mathcal{Q}$ is a *mid-rise* quantizer.

The *quantization error* $n$ is defined as

$$n \triangleq \mathcal{Q}(v) - v = w - v, \tag{2.46}$$

where

$$w \triangleq Q(v). \tag{2.47}$$

If the input to the scalar quantizer is a random variable $\mathrm{v}$, then the quantization error is also a random variable, denoted by $\mathrm{n}$. If the PDF of $\mathrm{v}$ is smooth, and if $L$ is large, then

$$\frac{\sigma_{\mathrm{v\,n}}}{\sigma_{\mathrm{n}}^2} \simeq 0, \tag{2.48}$$

see [130], and the quantization error has an approximately uniform PDF [94], which yields

$$\sigma_{\mathrm{n}}^2 = \frac{\Delta^2}{12}. \tag{2.49}$$

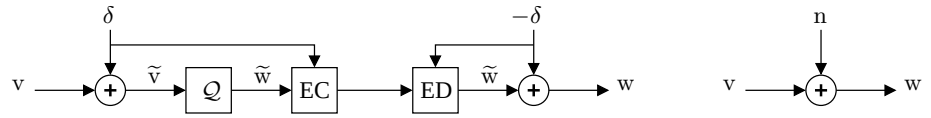



**Figure 2.1:** a) Subtractively dithered uniform quantization $\delta$ is a dither signal and EC and ED are, respectively, entropy encoder and entropy decoder. b) Equivalent model.

With the additional assumption that the PDF of $\mathrm{v}$ has bounded support, the variance of the quantization error can be well approximated, for large values of $L$, by

$$\sigma_{\mathrm{n}}^2 = c 2^{-2R_L} \sigma_{\mathrm{v}}^2, \tag{2.50}$$

see, e.g., [55], where

$$R_L \triangleq \log_2(L) \tag{2.51}$$

is the operational bit-rate of the quantizer in a *fixed-rate* quantization scenario, that is, if each value in $\mathbb{U}$ is encoded using $R_L$ bits. The multiplying factor $c$ in (2.50) depends on the PDF of $\mathrm{v}$.

The *signal-to-noise ratio* (SNR) associated with a scalar quantizer and its input is defined as

$$\gamma \triangleq \frac{\sigma_{\mathrm{v}}^2}{\sigma_{\mathrm{n}}^2} \tag{2.52}$$

Substituting (2.50) into (2.52), we can write

$$\gamma = c^{-1} 2^{2R_L}, \tag{2.53}$$

and

$$R_L = \frac{1}{2}\log_2(\gamma) + \frac{1}{2}\log_2(c) \tag{2.54}$$

### 2.4.2 Subtractively Dithered Uniform Scalar Quantization

A *subtractively dithered uniform scalar quantizer* (SDUSQ) is obtained by adding an i.i.d. dither signal $\{\delta(k)\} \sim \mathcal{U}[-\frac{\Delta}{2}, \frac{\Delta}{2}]$, statistically independent of $\{\mathrm{v}(k)\}$, to the input of the quantizer, and then subtracting $\{\delta(k)\}$ from the output [131–133]. This is shown in Fig. 2.1. The reconstructed output $\{\mathrm{w}(k)\}$ at time instant $k$ is given by

$$\mathrm{w}(k) = \mathcal{Q}(\mathrm{v}(k) + \delta(k)) - \delta(k), \quad \forall k \in \mathbb{Z}, \tag{2.55}$$

and the resulting quantization error

$$\mathrm{n}(k) = \mathrm{w}(k) - \mathrm{v}(k) = \mathrm{w}(k) + \delta(k) - \mathcal{Q}(\mathrm{v}(k) + \delta(k)), \quad \forall k \in \mathbb{Z}, \tag{2.56}$$

is i.i.d., uniformly distributed over $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$, and *independent of the input process* $\{\mathrm{v}(k)\}$ [132, 134]. The *asymptotic memoryless operational rate* of the SDUSQ is defined as

$$R_{\mathcal{Q}} \triangleq H(\tilde{\mathrm{w}}(k)|\delta(k)). \tag{2.57}$$

This quantity is independent of $k$ since $\tilde{\mathrm{w}}(k)$ and $\delta(k)$ are jointly stationary. $R_{\mathcal{Q}}$ corresponds to the rate (in bits/sample) achieved by a memoryless entropy coder acting on consecutive non-overlapping blocks of $N$ quantized values, when $N \to \infty$, supposing that the entropy coder assigns the word-length for the $\ell$-th block, $\ell \in \mathbb{Z}$, as an integer-valued approximation of $\sum_{k=\ell}^{\ell+N-1} H(\tilde{\mathrm{w}}(k)|\delta(k))$.

It was shown in [126] that

$$H(\tilde{\mathrm{w}}|\delta) = I(\mathrm{v};\mathrm{w}). \tag{2.58}$$

Also, from Lemma 4.10 in Section 4.8.2 of the current thesis,

$$I(\mathrm{v};\mathrm{w}) \leq \frac{1}{2}\log_2(\gamma+1) + 0.254 \text{ [bits/sample]}, \tag{2.59}$$

where equality holds if and only if $\mathrm{v}$ is Gaussian. Substituting (2.59) into (2.58) and (2.57),

$$R_{\mathcal{Q}} \leq \frac{1}{2}\log_2(\gamma+1) + 0.254 \text{ [bits/sample]}, \tag{2.60}$$

with equality if and only if $\mathrm{v}$ is Gaussian.

# Chapter 3

# WCMSE-Optimal Filters for a Given Quantizer SNR

*It's not easy taking my problems one at a time when they refuse to get in line.*
*Ashleigh Brilliant, British Cartoonist.*

*There are no small problems. Problems that appear small are large problems that are not understood.*
*Santiago Ramón y Cajal, Spanish neuroanatomist.*

*Divide and rule, a sound motto. Unite and lead, a better one.*
*Johann Wolfgang von Goethe, German poet, novelist and philosopher.*

## 3.1 Introduction

In this chapter we derive the optimal performance and frequency responses of the filters of full-band scalar quantization schemes, subject to a constraint on the SNR of the scalar quantizer. These results are an extension of earlier work by the author and colleagues, recently published in [118].

The general architecture for full-band scalar quantization schemes consists of a scalar quantizer and a set of linear filters around it, as shown in Fig. 3.1. We call this architecture a *feedback quantizer* (FQ). Well known examples of FQs include $\Delta$-Modulators, DPCM converters [55] and Sigma-Delta modulators [78]. The latter schemes have been very successfully applied in a number of areas, including audio compression [46, 55], oversampled A/D conversion [56, 78], sub-band coding [61], digital image half-toning [48, 76, 77], power conversion [135], and control over networks [136].

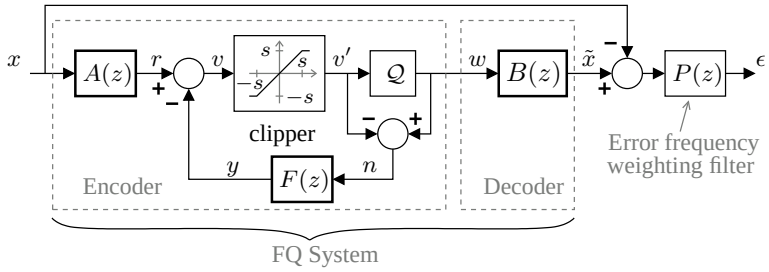



**Figure 3.1:** Feedback Quantization system and frequency weighting filter.

In this scheme, $\mathcal{Q}$ may take the form of a non-uniform or a uniform scalar quantizer [71], the latter being either dithered [1] or un-dithered [85].

The filters $A(z)$ and $B(z)$ in an FQ system allow one to exploit the predictability of the input signal so as to reduce the variance of $\{v(k)\}_{k\in\mathbb{Z}}$. When compared with simple PCM conversion, this flexibility allows one to use a scalar quantizer with a smaller quantization step. The error-feedback filter $F(z)$ opens the possibility of spectrally shaping the effect of quantization errors on the output. In this way, one can allocate more of the quantization noise in the frequency bands where it is less harmful from a user's point of view. Accordingly, it is convenient to use a frequency weighted error criterion, via an *error frequency weighting filter* $P(z)$, and to focus on the *frequency weighted error* $\epsilon$.

For the sake of generality, we consider the possible use of a clipper before $\mathcal{Q}$. This device limits the value of the quantizer input signal $v'$ so that

$$v' = \begin{cases} v, & \text{if } |v| \leq s, \\ \frac{v}{|v|}s, & \text{if } |v| > s, \end{cases}$$

where $s > 0$ is the *saturation threshold* of the clipper. This clipping technique, which is equivalent to the one proposed in [56], can be used to keep $\mathcal{Q}$ from overloading, which is helpful in reducing limit-cycle oscillations (idle tones) in an FQ with high order filters. On the other hand, if we chose $s$ to be sufficiently large, then $v' = v$, and the clipper has no effect on the system.

If the characteristics of $\mathcal{Q}$ and the spectral properties of the input signal $x$ are known, then the design of an FQ converter that minimizes the variance of $\epsilon$ amounts to choosing the filters $A(z)$, $B(z)$ and $F(z)$.

In this chapter, we will characterize the performance and associated filters of optimal feedback quantizers, under different architectural limitations. (See Section 1.1.3). For this purpose, as in [79–83], we model the scalar quantizer as a linear device that introduces additive white noise whose variance is proportional to that of the signal being quantized. The main results are:

[1] In this case, the block $\mathcal{Q}$ in Fig. 3.1 represents the scalar quantizer including the dither signals.

1. We derive equations that relate the minimum achievable frequency weighted WCMSE to the *signal-to-noise ratio* (SNR) of $\mathcal{Q}$, for any subset of the filters $A(z)$, $B(z)$ and $F(z)$ being given and fixed. Each possible subset of filters being fixed gives rise to a different optimization problem. These optimization problems were listed in Table 1.1 (page 18), for MSE as the distortion metric. We solve these problems using the WCMSE as the distortion metric, which includes the MSE problems as special cases. We derive equations that characterize the optimal filters for each case. Within the scope of validity of the Linear Model, our results can be applied to quantizers having any given number of quantization levels, and to almost arbitrary input spectra and frequency weighting criteria.

2. We show that, within our model, the frequency weighted MSE in an optimal FQ where the SNR of $\mathcal{Q}$ is fixed, decreases exponentially with oversampling ratio $\lambda$. From this result it follows that, if $\mathcal{Q}$ is an entropy coded subtractively dithered scalar quantizer with sufficient quantization levels to avoid overload (or clipping), then, for a fixed operational bit-rate,

$$MSE = \mathcal{O}(2^{-1.746\lambda}),$$

as $\lambda \to \infty$. We also derive an extension of this result for the case of subtractively dithered scalar quantization with a (finite) number of quantization levels that is insufficient to avoid quantizer overload. This covers situations in which the source samples, $\mathrm{x}(i)$, have a PDF with unbounded support, provided that the moments $\mu_n^{(i)} \triangleq \mathrm{E}\left[\mathrm{x}(i)^n\right]$ can be bounded as

$$\left|\mu_n^{(i)}\right| \leq \frac{1}{2}(n!)H^{n-2}\left|\mu_2^{(i)}\right|, \quad \forall n \geq 2, \forall i \in \mathbb{Z}^+, \tag{3.1}$$

for some finite scalar $H$. We note that this requirement is satisfied by most PDFs of practical or theoretical interest, and in particular, by uniform, Gaussian and Laplacian PDFs. For these cases, we show that if $\mathcal{Q}$ is a subtractively dithered scalar quantizer with $N$ levels, then the MSE can be made to decay as

$$MSE = \mathcal{O}(\mathrm{e}^{-c_0\lambda^{1/3}}),$$

when $\lambda \to \infty$, where $c_0 \triangleq 4N^{2/3}$. In order to achieve this asymptotic decay rate, it is necessary to balance the variance of clipping and granular errors in the quantizer, for each oversampling ratio, by carefully adjusting the loading factor (defined as half the input dynamic range of the quantizer divided by the standard deviation of its zero-mean input signal) at which $\mathcal{Q}$ operates. To the best of the author's knowledge, this is the only result in the literature combining overloading quantization and oversampling. It also seems to be the first decay rate bound for the MSE of oversampled quantization that holds for sources with infinite support.

The contents of this chapter are organized as follows: In Section 3.2, we present our analysis model for PRFQ converters. The different optimization problems arising from different architectural constraints are solved in sections 3.3–3.9. The case of oversampled FQ is analyzed in Section 3.12. Section 3.9.2 discusses the relationship to previous results and highlights the importance of taking account of fed back quantization noise. Section 3.11 presents a simulation example, and Section 3.13 summarizes the main results of the chapter.

## 3.2 Analysis Model and Assumptions

In this section we discuss some of the main aspects of feedback quantization. We also describe the analysis model and the constraints to be considered.

### 3.2.1 Feedback Quantizer Equations

We begin by presenting the equations that describe the behaviour of the FQ shown in Fig. 3.1.

#### Quantization and Clipping Errors

From Fig. 3.1, the quantization error $n$ is given by

$$n(k) \triangleq w(k) - v'(k). \tag{3.2}$$

Every practical scalar quantizer has an associated constant $V > 0$ such that, if $|v'| > V$, then $\mathcal{Q}$ is said to be *overloaded*. When the quantizer is not overloaded, then $n(k)$ is said to consist of only *granular* quantization error, namely $\varrho(k)$, which can be bounded as $|\varrho(k)| \leq \varrho_{max}$, $\forall v'(k) \in \mathbb{R}$, for some $0 < \varrho_{max} < 2V$ (see, e.g., [71]). For example, if $\mathcal{Q}$ is a symmetric, uniform, non-dithered quantizer with $N$ levels and quantization interval $\Delta$, then one needs $V \leq N\Delta/2$ in order to obtain $\varrho_{max} = \frac{\Delta}{2}$.

In general, we can write

$$n(k) = \varrho(k) + \tau(k), \tag{3.3}$$

where

$$\tau(k) \triangleq v(k)' - \frac{v(k)'}{|v(k)'|} \min\{V, |v'(k)|\}$$

is the *overload* error. Clearly overload errors are bounded as $|\tau(k)| < |v'(k)| \leq |v(k)|$, but they cannot be bounded by a constant unless $v'$ is bounded.

As outlined in the introduction, the clipper in Fig. 3.1 can be used to keep $\mathcal{Q}$ from overloading. For simplicity, we will consider only two possibilities, namely, that $s = V$, or else $s = \infty$. The former choice guarantees that $\mathcal{Q}$ does not overload, since the *clipping error*, defined as

$$\vartheta(k) \triangleq v'(k) - v(k), \quad \forall k \in \mathbb{Z}, \tag{3.4}$$

takes place instead. More precisely, if $s = \infty$ we have that

$$\vartheta(k) = 0, \quad \text{and} \tag{3.5}$$

$$\tau(k) = v(k) - \frac{v(k)}{|v(k)|} \min\{V, |v(k)|\}. \tag{3.6}$$

If, instead, $s = V$, then the latter revert to

$$\vartheta(k) = v(k) - \frac{v(k)}{|v(k)|} \min\{V, |v(k)|\} \quad \text{and} \tag{3.7}$$

$$\tau(k) = 0. \tag{3.8}$$

A key point in using clipping is that, unlike overload errors, clipping errors are not fed back into $\mathcal{Q}$ through $F(z)$. This helps to avoid large limit-cycle oscillations arising from the overload of $\mathcal{Q}$, see [56]. Since such oscillations are not part of the analysis model we will use, their occurrence could increase the frequency weighted WCMSE significantly above the value predicted by the model.

Using the above definitions, and from Fig. 3.1, we can write

$$w(k) = v(k) + n(k) + \vartheta(k), \tag{3.9}$$

which reveals that $w$ differs from $v$ by the sum of the quantization and clipping errors.

### Transfer Functions

From Fig. 3.1 and (3.9) we have that

$$v = A(z)x - F(z)n, \tag{3.10a}$$

$$\tilde{x} = B(z)A(z)x + B(z)[1 - F(z)]n + B(z)\vartheta, \tag{3.10b}$$

$$\epsilon = P(z)B(z)[1 - F(z)]n + P(z)B(z)\vartheta. \tag{3.10c}$$

Notice that these equations are exact and require no assumptions on the signals involved. From (3.10b) one can see that $A(z)B(z)$ corresponds to the *signal transfer function* (STF), from $x$ to $\tilde{x}$, of the converter. Similarly, the product $B(z)[1-F(z)]$ is the transfer function for quantization errors. It is usually referred to as the *noise transfer function* (NTF) of the converter[2]. The term $[1 - F(z)]$ will play a crucial role in the derivation of the optimal filters in the subsequent sections.

[2]In noise-shaping and $\Sigma\Delta$ literature, where $B(z)$ is typically a unit gain, the term NTF is normally used for $1 - F(z)$.

**Stability**

We say that a PRFQ is *Bounded-Input-Bounded Output* (BIBO) stable if and only if for any input sequence $x$ satisfying $\|x\|_\infty \leq x_{max} < \infty$ all the signals in the converter are bounded.

If $s = V$, or if $\mathcal{Q}$ has infinitely many quantization levels, then $|n| \leq \varrho_{max}$, $\forall k \in \mathbb{Z}$, and thus all the other signals in the converter are bounded. On the other hand, if $s = \infty$, then, from Fig. 3.1, $v$ can be written as

$$v = \frac{A(z)}{1 - F(z)} x - \frac{F(z)}{1 - F(z)} w. \tag{3.11}$$

If the quantizer has a finite number of quantization levels, then $w$ is bounded. If $F(z)$ is stable and $1 - F(z)$ is minimum-phase, then it follows from (3.11) that $v$ is bounded. This in turn guarantees that $n$ and all the other signals in the converter are bounded (see (3.2) and (3.10)). Summarizing, if all the filters in Fig. 3.1 are stable, and if $1 - F(z)$ has no zeros on or outside the unit circle, then the resulting PRFQ is BIBO stable.

In addition, if $A(z)$ and $F(z)$ are stable, then the $\ell_\infty$ norm of their impulse responses, namely $A_\infty$ and $F_\infty$, are bounded. Thus, if there exists a bounded $x_{max} > 0$ such that $|x(k)| \leq x_{max} < \infty$, $\forall k \in \mathbb{Z}$, then a sufficient condition to ensure $\tau(k) = \vartheta(k) = 0$, $\forall k \in \mathbb{Z}$, is that $V \geq V_{min} < \infty$, where

$$V_{min} \triangleq A_\infty x_{max} + F_\infty \varrho_{max}. \tag{3.12}$$

Thus, for a uniform quantizer with quantization interval $\Delta$, it suffices to have $V_{min}/\Delta$ or more quantization levels in order to avoid clipping or overload errors. The latter condition provides a "worst-case" stability criterion, which has been considered, e.g., in [96, 137, 138].

### 3.2.2 Assumptions

The assumptions associated with our FQ model are described next.

**Input Spectrum and Frequency Weighting**

The error weighting filter $P(z)$ in Fig. 3.1 models the impact that reconstruction errors have at each frequency. This "performance assessment" filter is application dependent, and is assumed to be stable and given. The input signal $\{x(k)\}_{k\in\mathbb{Z}}$ is a zero-mean w.s.s. stochastic process[3] with known PSD $S_x(\omega) = \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^2$ and finite power, i.e., $\|\Omega_{\mathrm{x}}\|^2 < \infty$. In order to simplify our subsequent analysis, we shall further restrict $\Omega_{\mathrm{x}}$ and $P(z)$ to satisfy the following:

[3] This excludes, for example, non-zero mean random signals, sinusoids, or constant inputs from the analysis.

**Assumption 3.1.** The product $|\Omega_{\mathrm{x}}P|$ is a piece-wise differentiable function having, at most, a finite number of discontinuities and satisfying $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| < \infty,\ \forall\omega \in [-\pi,\pi]$. In addition, $|\Omega_{\mathrm{x}}P|$ is such that one[4] of the following conditions holds:

i) There exists a constant $g_{min} > 0$ such that $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| > g_{min}$, for all $\omega \in [-\pi,\pi]$, or

ii) $\exists\omega \in [-\pi,\pi]$ such that $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| = 0$. Furthermore, if $\{\Gamma_i\}$ denotes the set of non-contiguous and non-overlapping intervals in $[-\pi,\pi]$ such that $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| = 0 \Leftrightarrow \omega \in \bigcup_i \Gamma_i$, then, for every $i$, $\exists\zeta_i \in \Gamma_i$ such that $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$ is $\mathcal{O}(\omega-\zeta_i)$ as $\omega \to \zeta_i$. ▲

We note that the above is a rather weak constraint, since conditions i) and ii) include almost any product $|\Omega_{\mathrm{x}}P|$ of practical or theoretical interest. In particular, condition i) covers all the cases where the product $\Omega_{\mathrm{x}}(z)P(z)$ has no zeros on the unit circle. In turn, condition ii) is satisfied if $P\Omega_{\mathrm{x}}$ is zero over any interval on $[-\pi,\pi]$ having non-zero measure, or if $\Omega_{\mathrm{x}}(z)P(z)$ is rational and has zeros on the unit circle.

### The Quantizer

We shall focus our analysis on the effect that granular quantization errors have on the *frequency-weighted* WCMSE, (FW-WCMSE). For their effect to closely represent the actual FW-WCMSE, we need to assume the following:

**Assumption 3.2.** The variances of overload and clipping errors are negligible, i.e.,

$$\sigma_\tau^2 \ll \sigma_n^2, \quad \text{if } s=\infty\text{, or} \tag{3.13a}$$

$$\sigma_\vartheta^2 \ll \sigma_n^2, \quad \text{if } s=V. \tag{3.13b}$$

▲

In addition, and as stated in the introduction, we will adopt an additive white noise model for $n$. This model is widely used for the analysis and design of data converters (see, e.g., [43, 46, 55, 56, 61, 78–83, 87, 90, 92]). It is usually described as follows:

**Assumption 3.3.** The sequence of quantization noise $\{n(k)\}_{k\in\mathbb{Z}}$ is a zero-mean w.s.s. random process, uncorrelated with the input of the PRFQ, and having constant PSD

$$S_n(\omega) = \sigma_n^2, \quad \forall\omega \in [-\pi,\pi],$$

where $\sigma_n^2$ is the variance of $\{n(k)\}_{k\in\mathbb{Z}}$. ▲

[4] Notice that conditions i) and ii) cannot be met simultaneously.

The above additive white noise model, although not exact, is, in general, a good approximation when a signal with a smooth *probability density function* (PDF) is quantized with many levels and negligible overload (in the sense of Assumption 3.2), see, e.g., [78]. The model can be made exact, *even for few quantization levels*, by utilizing a uniform scalar quantizer with either subtractive or non-subtractive dither[5], provided quantizer overload does not occur, see [85]. As discussed before, one way to achieve this is to use a quantizer with a sufficiently large number of quantization levels, so as to satisfy (3.12). In this case, if the quantization interval is $\Delta$ and the dither sequence $\delta$ whitens $n$, makes $n$ uncorrelated to $x$ when $\mathcal{Q}$ is not overloaded, and is bounded as $|\delta(k)| \leq \delta_{max}$, $\forall k \in \mathbb{Z}$, then any number of levels greater than or equal to $(V_{min} + 2\delta_{max})/\Delta$ will make Assumption 3.3 hold exactly. If a smaller number of quantization levels are employed so that $V < V_{min}$, then the use of dither with the same characteristics as before, together with clipping (i.e., setting $s = V$), will also make $n$ satisfy Assumption 3.3 exactly.

Assumption 3.3 allows one to write the variance of $\{v(k)\}_{k\in\mathbb{Z}}$ as

$$\sigma_v^2 = \|A\Omega_x\|^2 + \sigma_n^2\|F\|^2, \tag{3.14}$$

see Fig. 3.1. This equation describes the effect of $\sigma_n^2$ on $\sigma_v^2$ through the feedback path. However, if the scalar quantizer has a finite and fixed number of quantization levels, then another link between these two variances needs to be considered. In order to model this relationship, we will use the fixed signal-to-noise ratio model employed in, e.g., [79–82, 87]:

**Assumption 3.4.** *For a fixed number of quantization levels, the variance of quantization errors is proportional to the variance of the signal being quantized, i.e.,*

$$\gamma \triangleq \frac{\sigma_v^2}{\sigma_n^2}. \tag{3.15}$$

*is fixed.* ▲

If no clipping is used (i.e., if $s = \infty$), then $\gamma$ corresponds exactly to the SNR of $\mathcal{Q}$. If $s = V$, then $\gamma$ is a good a approximation to the SNR of $\mathcal{Q}$ when (3.13b) in Assumption 3.2 holds.

In our model, $\gamma$ is assumed fixed and given. Strictly speaking, as already mentioned in Section 2.4, $\gamma$ depends on the PDF of $\{v(k)\}_{k\in\mathbb{Z}}$, on the number of quantization levels of $\mathcal{Q}$, and on how quantization thresholds and levels are distributed along the dynamic range of $\mathcal{Q}$. In practice, for a given number of quantization levels, $\gamma$ should be chosen such that the dynamic range of $\mathcal{Q}$ is used efficiently, whilst ensuring a low probability of quantizer overload or clipping. For example, for the often cited uniform quantizer with $N$ levels and loading factor[6] equal to $4$ we obtain $\gamma = \frac{3}{16}N^2$ (assuming that $\{n(k)\}_{k\in\mathbb{Z}}$

[5]Here and in the sequel we assume the dither is such that $n$ is white and uncorrelated with $x$ when $\mathcal{Q}$ is not overloaded.

[6]The loading factor corresponds to the ratio between half the dynamic range of $\mathcal{Q}$ and the standard deviation of its input.

has a uniform PDF and negligible overload errors). We note that, for large $N$, and provided the signal being quantized has bounded support, a quadratic relationship between $N$ and $\gamma$ holds for most types of scalar quantizers (see, e.g., [71, 130]). This is indeed the well known rule of "6 [dB] reduction of quantization noise variance per additional bit of quantizer resolution" [55, 71].

In the sequel, we refer to the model of quantization errors determined by Assumptions 3.2, 3.3 and 3.4 as the **Linear Model**. Summarizing, the Linear Model is exact if the FQ uses a dithered quantizer having enough quantization levels to avoid overload. If only an insufficient number of quantization levels is available, but dither is used jointly with clipping, then the model is exact for predicting the effects of granular quantization errors, and is a good approximation for predicting the total frequency weighted WCMSE if Assumption 3.2 also holds. If the scalar quantizer is un-dithered, has a small quantization interval (relative to $\sigma_{v'}$) and enough quantization levels to avoid overload, then the Linear Model can be expected to yield a good approximation to the total frequency weighted WCMSE. Perhaps surprisingly, the Linear Model turns out to predict with remarkable accuracy the WCMSE of an optimal PRFQ when few quantization levels and clipping are used with a loading factor big enough to satisfy Assumption 3.2, even without dither, and even for a 1-bit quantizer. This is evident in the simulation results presented later in Section 3.11.

### 3.2.3 Optimization Constraints

The filters $A(z)$, $B(z)$ and $F(z)$ in Fig. 3.1 are design choices. We shall restrict the search for the optimal filters to those satisfying the following constraint:

**Constraint 3.1.**

1. $A(z)$ and $B(z)$ are stable.
2. $F(z)$ is stable and strictly causal (i.e., $\lim_{z\to\infty} F(z) = 0$). ▲

As discussed in Section 3.2.1, the stability constraints on $A(z)$, $B(z)$ and $F(z)$ are a necessary condition for the converter to be BIBO stable. The additional requirement on $F(z)$, namely strict causality, is needed for the feedback loop in Fig. 3.1 to be well defined (see, e.g., [78, Chap. 4]). Notice that we will not, a priori, require $1 - F(z)$ to have zeros only inside the open unit disk. Instead, we will show that the latter property arises naturally from the solution of the design optimization problem.

An additional constraint on $F(z)$ arises from the value of $\gamma$, as explained next. The ratio between the variances of $v$ and $n$ *imposed by the feedback* can be obtained by dividing (3.14) by $\sigma_n^2$, yielding

$$\frac{\sigma_v^2}{\sigma_n^2} = \frac{\|A\Omega_{\mathrm{x}}\|^2}{\sigma_n^2} + \|F\|^2. \tag{3.16}$$

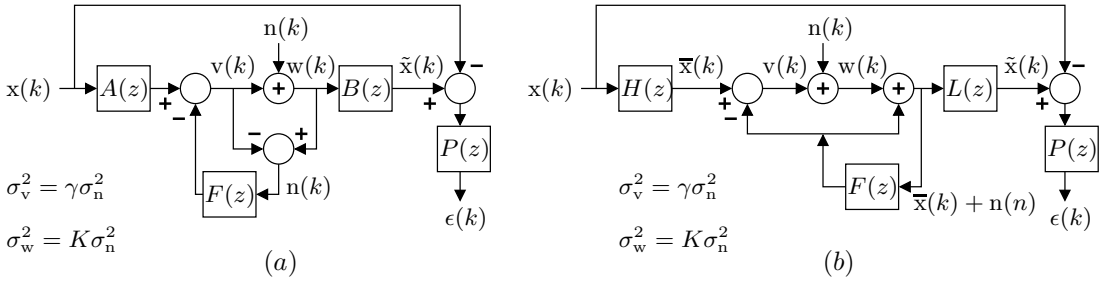



**Figure 3.2:** Equivalent analysis models

One can see from the above that if $\|F\|^2 > \gamma$, then *any* pre-filter or scaling of the quantization intervals of $\mathcal{Q}$ will yield $\sigma_v^2 > \gamma\sigma_n^2$, thus making large overload (or clipping) inevitable. This would increase overall distortion, and if no clipping is used, may lead to large limit-cycle oscillations. We thus conclude that the use of feedback imposes the following constraint:

**Constraint 3.2.**

$$\|F\|^2 < \gamma.$$

▲

### 3.2.4 Analysis Model

Under the Linear Model originating from assumptions 3.2, 3.3 and 3.4, the feedback quantizer of Fig. 3.1 can be analyzed using the system shown in Fig. 3.2-(a). In this figure, $\{\mathrm{n}(k)\}$ is a white w.s.s. process uncorrelated with $\{\mathrm{x}(k)\}$, and whose variance, $\sigma_{\mathrm{n}}^2$, obeys the SNR assumption (3.15). The variance of $\{\mathrm{n}(k)\}$ can be found by substituting (3.15) into (3.16). This gives

$$\sigma_n^2 = \frac{\|A\Omega_x\|^2}{\gamma - \|F\|^2}. \tag{3.17}$$

Since $F(z)$ is strictly causal, it satisfies

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} F(\mathrm{e}^{j\omega})d\omega = 0. \tag{3.18}$$

Thus, we have that

$$\|F\|^2 = \|1 - F\|^2 - 1. \tag{3.19}$$

Substitution of (3.19) into (3.17) yields

$$\sigma_{\mathrm{n}}^2 = \frac{\|\Omega_{\mathrm{x}}A\|^2}{\gamma + 1 - \|1 - F\|^2} \tag{3.20}$$

Since, in the Linear Model, quantization errors are uncorrelated with the source, the frequency-weighted WCMSE (see (1.11) on page 26) is given by

$$D_{a,b}(\mathrm{x},\mathrm{y}) = a\sigma_{\mathrm{n}}^2\|PB(1-F)\|^2 + b\|(AB-1)\Omega_{\mathrm{x}}P\|^2 \tag{3.21a}$$

$$= a\frac{\|\Omega_{\mathrm{x}}A\|^2\|PB(1-F)\|^2}{\gamma+1-\|1-F\|^2} + b\|(AB-1)\Omega_{\mathrm{x}}P\|^2 \tag{3.21b}$$

Upon defining

$$f(\mathrm{e}^{j\omega}) \triangleq \left|1-F(\mathrm{e}^{j\omega})\right|, \quad \forall\omega\in[-\pi,\pi] \tag{3.22a}$$

$$K \triangleq \gamma+1 \tag{3.22b}$$

we can re-write (3.21b) more compactly as

$$D_{a,b}(\mathrm{x},\mathrm{y}) = a\frac{\|\Omega_{\mathrm{x}}A\|^2\|PBf\|^2}{K-\|f\|^2} + b\|(AB-1)\Omega_{\mathrm{x}}P\|^2 \tag{3.23}$$

In the subsequent analysis, it will also be useful to consider the equivalent structure shown in Fig. 3.2-(b). This scheme is equivalent to the one depicted in Fig. 3.2-(a) if and only if

$$H(z) = \frac{A(z)}{1-F(z)}, \tag{3.24a}$$

$$L(z) = [1-F(z)]\,B(z). \tag{3.24b}$$

The equivalent expression for the frequency weighted WCMSE for the system in Fig. 3.2-(b) can be readily obtained upon substituting (3.24) into (3.23), yielding:

$$D_{a,b}(\mathrm{x},\mathrm{y}) = a\frac{\|PL\|^2\|\Omega_{\mathrm{x}}Hf\|^2}{K-\|f\|^2} + b\|(HL-1)\Omega_{\mathrm{x}}P\|^2 \tag{3.25}$$

Comparison of (3.25) with (3.23) reveals the duality between the two schemes shown in Fig. 3.2.

In sections 3.3 to 3.9, we characterize the frequency response of the filters that minimize $D_{a,b}(\mathrm{x},\mathrm{y})$ subject to the constraint that $\gamma$ is fixed, and subject to constraints 3.1 and 3.2. We consider different scenarios of architectural limitations, as described in Section 1.1.3. In each scenario, a different subset of the filters of the system in Fig. 3.2-(a) (or of the system in Fig. 3.2-(b)) is considered to be given and fixed. Thus, each scenario generates a different optimization problem. We begin in Section 3.3 with the most restrictive scenarios, in which only one degree of freedom is available. We finish in Section 3.9 by solving the problem in which the three filters that minimize the FWCMSE have to be found, i.e. where there are are no architectural limitations.

## 3.3 $F(z)$ and $A(z)$ (or $B(z)$) Given

### 3.3.1 $F(z)$ and $A(z)$ Given

If $A(z)$ and $F(z)$ are given and fixed, minimization of the frequency weighted WCMSE reduces to the following:

**Optimization Problem 3.1.** *For a given $K > 0$, frequency response $A(z)$ and the frequency response magnitudes $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|$, $\left|P(\mathrm{e}^{j\omega})\right|$, $f(\mathrm{e}^{j\omega})$, find the filter $B(z)$ that minimizes*

$$D_{a,b}(\mathrm{x},\mathrm{y}) = a\frac{\|\Omega_{\mathrm{x}}A\|^2\|PBf\|^2}{K-\|f\|^2} + b\|(AB-1)\Omega_{\mathrm{x}}P\|^2. \tag{3.26}$$

▲

The answer to this problem is given in the following:

**Theorem 3.1.** *The solution to Optimization Problem 3.1 satisfies*

$$B(\mathrm{e}^{j\omega}) = \frac{b\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^2 A(\mathrm{e}^{j\omega})^*}{a\sigma_n^2 f(\mathrm{e}^{j\omega})^2 + b\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})A(\mathrm{e}^{j\omega})\right|^2}, \quad \textit{a.e. on } [-\pi,\pi]\setminus\mathcal{N}_P, \tag{3.27}$$

*where $\sigma_n^2$ is given by* (3.20)*. The frequency response of the solution, $B(\mathrm{e}^{j\omega})$, can take any arbitrary (bounded) value for all $\omega \in \mathcal{N}_P$. The minimum $D_{a,b}$, achieved with $B(z)$ as in* (3.27)*, is*

$$\min_{B(z)} D_{a,b}(\mathrm{x},\mathrm{y}) = ab\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2 f(\mathrm{e}^{j\omega})^2}{af(\mathrm{e}^{j\omega})^2 + b\frac{K-\|f\|^2}{\|\Omega_{\mathrm{x}}A\|^2}\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})A(\mathrm{e}^{j\omega})\right|^2} d\omega \tag{3.28}$$

▲

*Proof.* Define

$$W(z) \triangleq A(z)B(z). \tag{3.29}$$

Noting that the optimal $B(\mathrm{e}^{j\omega})$ must clearly be bounded a.e. on $[-\pi,\pi]$, we can write

$$B(\mathrm{e}^{j\omega}) = A(\mathrm{e}^{j\omega})^{\sim 1}W(\mathrm{e}^{j\omega}) \tag{3.30}$$

Substitution into (3.21a) yields

$$D_{a,b}(\mathrm{x},\mathrm{y}) = a\sigma_n^2\|PfA^{\sim 1}W\|^2 + b\|(W-1)\Omega_{\mathrm{x}}P\|^2$$

Applying Lemma 3.16 (page 104) to this equation, the transfer function $W(z)$ that minimizes $D_{a,b}(\mathrm{x},\mathrm{y})$ is found to satisfy

$$W(\mathrm{e}^{j\omega}) = \frac{b\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2\left|P(\mathrm{e}^{j\omega})\right|^2}{a\sigma_n^2\left|P(\mathrm{e}^{j\omega})\right|^2 f(\mathrm{e}^{j\omega})^2\left(\left|A(\mathrm{e}^{j\omega})\right|^{\sim 1}\right)^2 + b\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2\left|P(\mathrm{e}^{j\omega})\right|^2} \tag{3.31}$$

$$= \frac{b\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2\left|A(\mathrm{e}^{j\omega})\right|^2}{a\sigma_n^2 f(\mathrm{e}^{j\omega})^2 + b\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2\left|A(\mathrm{e}^{j\omega})\right|^2}, \quad \text{a.e. on } [-\pi,\pi]. \tag{3.32}$$

Notice that (3.32) follows by multiplying both sides in (3.31) by $\left|A(\mathrm{e}^{j\omega})\right|^2 \left(\left|A(\mathrm{e}^{j\omega})\right|^{\sim 1}\right)^2$, and by noting, from (3.30), that $W(\mathrm{e}^{j\omega})\left|A(\mathrm{e}^{j\omega})\right|^2 \left(\left|A(\mathrm{e}^{j\omega})\right|^{\sim 1}\right)^2 = W(\mathrm{e}^{j\omega})$, $\forall \omega \in [-\pi, \pi]$. Substitution of (3.32) into (3.30), together with the fact that $\left|A(\mathrm{e}^{j\omega})\right|^2 A(\mathrm{e}^{j\omega})^{\sim 1} = A(\mathrm{e}^{j\omega})^*$, yield (3.27). Substitution of (3.27) into (3.26) yields (3.28). This completes the proof. □

### 3.3.2 $F(z)$ and $B(z)$ Given

If $F(z)$ and $B(z)$ are given then the minimization of the WCMSE can be stated as the following optimization problem:

**Optimization Problem 3.2.** *For a given $K > 0$, frequency response $B(z)$ and the frequency response magnitudes $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|$, $\left|P(\mathrm{e}^{j\omega})\right|$, $f(\mathrm{e}^{j\omega})$, find the filter $A(z)$ that minimizes*

$$D_{a,b}(\mathrm{x},\mathrm{y}) = a\frac{\|\Omega_{\mathrm{x}}A\|^2\|PBf\|^2}{K-\|f\|^2} + b\|(AB-1)\Omega_{\mathrm{x}}P\|^2. \tag{3.33}$$

▲

The answer to this problem is given in the following:

**Theorem 3.2.** *The solution to Optimization Problem 3.2 satisfies*

$$A(\mathrm{e}^{j\omega}) = \frac{b\left|P(\mathrm{e}^{j\omega})\right|^2 B(\mathrm{e}^{j\omega})^*}{aV + b\left|P(\mathrm{e}^{j\omega})\right|^2 \left|B(\mathrm{e}^{j\omega})\right|^2}, \quad \textit{a.e. on } [-\pi,\pi]\setminus\mathcal{N}_{\Omega_{\mathrm{x}}}, \tag{3.34a}$$

*where*

$$V \triangleq \frac{\|PBf\|^2}{K-\|f\|^2}. \tag{3.34b}$$

*The frequency response of the solution, $A(\mathrm{e}^{j\omega})$, can take any arbitrary (bounded) value for all $\omega \in \mathcal{N}_{\Omega_{\mathrm{x}}}$. The minimum $D_{a,b}$, achieved with $A(z)$ as in* (3.34a)*, is*

$$\min_{A(z)} D_{a,b}(\mathrm{x},\mathrm{y}) = ab\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2}{a + b\frac{K-\|f\|^2}{\|PBf\|^2}\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2} d\omega \tag{3.35}$$

▲

*Proof.* Noting from (3.33) that the optimal $A(\mathrm{e}^{j\omega})$ must be bounded a.e. on $[-\pi,\pi]$, we can write

$$A(\mathrm{e}^{j\omega}) = B(\mathrm{e}^{j\omega})^{\sim 1}W(\mathrm{e}^{j\omega}), \quad \text{a.e. on } [-\pi,\pi], \tag{3.36}$$

where $W(z)$ is as defined in (3.29). Substitution of (3.36) and (3.29) into (3.23) yields

$$D_{a,b}(\mathrm{x},\mathrm{y}) = aV\|\Omega_{\mathrm{x}}B^{\sim 1}W\|^2 + b\|(W-1)\Omega_{\mathrm{x}}P\|^2, \tag{3.37}$$

where $V$ is as in (3.34b). It is clear from (3.37) that the optimal $W(\mathrm{e}^{j\omega})$ can take any arbitrary bounded values at all frequencies $\omega \in \mathcal{N}_{\Omega_\mathrm{x}}$. For all other frequencies, direct application of Lemma 3.16 (page 104) to (3.37) yields that the transfer function $W(z)$ that minimizes (3.37) satisfies

$$W(\mathrm{e}^{j\omega}) = \frac{b\left|\Omega_\mathrm{x}(\mathrm{e}^{j\omega})\right|^2\left|P(\mathrm{e}^{j\omega})\right|^2}{aV\left|\Omega_\mathrm{x}(\mathrm{e}^{j\omega})\right|^2\left(\left|B(\mathrm{e}^{j\omega})\right|^{\sim 1}\right)^2 + b\left|\Omega_\mathrm{x}(\mathrm{e}^{j\omega})\right|^2\left|P(\mathrm{e}^{j\omega})\right|^2} \tag{3.38}$$

$$= \frac{b\left|P(\mathrm{e}^{j\omega})\right|^2\left|B(\mathrm{e}^{j\omega})\right|^2}{aV + b\left|P(\mathrm{e}^{j\omega})\right|^2\left|B(\mathrm{e}^{j\omega})\right|^2}, \quad \text{a.e. on } [-\pi,\pi]\setminus\mathcal{N}_{\Omega_\mathrm{x}}. \tag{3.39}$$

We note that (3.39) follows by multiplying both sides in (3.38) by $\left|B(\mathrm{e}^{j\omega})\right|^2\left(\left|B(\mathrm{e}^{j\omega})\right|^{\sim 1}\right)^2$, and by noting, from (3.36), that $W(\mathrm{e}^{j\omega})\left|B(\mathrm{e}^{j\omega})\right|^2\left(\left|B(\mathrm{e}^{j\omega})\right|^{\sim 1}\right)^2 = W(\mathrm{e}^{j\omega})$, $\forall\omega \in [-\pi,\pi]$. Substitution of (3.39) into (3.36), together with the fact that $\left|B(\mathrm{e}^{j\omega})\right|^2 B(\mathrm{e}^{j\omega})^{\sim 1} = B(\mathrm{e}^{j\omega})^*$, yield (3.34a). Substitution of (3.34a) into (3.33) yields (3.35). This completes the proof. □

## 3.4 $F(z)$ and the Signal Transfer Function Given

If the signal transfer function $A(z)B(z)$ is set equal to a given transfer function $W(z)$, then the minimization of the WCMSE reduces to the following:

**Optimization Problem 3.3.** *For a given $K > 1$ frequency responses $\Omega_\mathrm{x}(\mathrm{e}^{j\omega})$, $P(\mathrm{e}^{j\omega})$, frequency response magnitude $f(\mathrm{e}^{j\omega})$, and transfer function $W(z)$, find the filters $A(z)$, $B(z)$ that*

$$\textit{minimize: } D_{a,b}(\mathrm{x},\mathrm{y}) = a\frac{\|\Omega_\mathrm{x}A\|^2\|PBf\|}{K-\|f\|^2} + b\|(AB-1)\Omega_\mathrm{x}P\|^2 \tag{3.40a}$$

$$\textit{subject to: } A(z)B(z) = W(z) \tag{3.40b}$$

The solution to this problem is provided by the theorem below.

**Theorem 3.3.** *The filters $A(z)$ and $B(z)$ that solve Optimization Problem 3.3 are completely characterized by the following equations:*

$$\left|A(\mathrm{e}^{j\omega})\right| = \kappa\sqrt{\left|P(\mathrm{e}^{j\omega})\right|\left|\Omega_\mathrm{x}(\mathrm{e}^{j\omega})\right|^{\sim 1} f(\mathrm{e}^{j\omega})\left|W(\mathrm{e}^{j\omega})\right|}, \quad \textit{a.e. on } [-\pi,\pi], \tag{3.41a}$$

$$\left|B(\mathrm{e}^{j\omega})\right| = \frac{1}{\kappa}\sqrt{\left|P(\mathrm{e}^{j\omega})\right|^{\sim 1}\left|\Omega_\mathrm{x}(\mathrm{e}^{j\omega})\right| f(\mathrm{e}^{j\omega})^{\sim 1}\left|W(\mathrm{e}^{j\omega})\right|}, \quad \textit{a.e. on } [-\pi,\pi], \tag{3.41b}$$

*where $\kappa > 0$ is an arbitrary real constant. The minimum of* (3.40a) *under the constraint* (3.40b)*, which is achieved by filters $A(z)$ and $B(z)$ satisfying* (3.41)*, is*

$$\min_{A(z)B(z)=W(z)} D_{a,b}(\mathrm{x},\mathrm{y}) = \sigma^2_{\epsilon\, inf|F} \triangleq a\frac{\langle f, |\Omega_\mathrm{x}P|\,|W|\rangle^2}{K-\|f\|^2} + b\|(W-1)\Omega_\mathrm{x}P\|^2 \tag{3.42}$$

▲

*Proof.* Denote the numerator of the first term on the right side term of (3.40a) as

$$N \triangleq \|\Omega_x A\|^2 \|(1-F)PB\|^2.$$

Applying Cauchy-Schwartz inequality we get

$$N \geq \langle |\Omega_x A|, |(1-F)PB| \rangle^2 = \langle |\Omega_x P|\,|AB|, |1-F| \rangle^2 = \langle |W|\,|\Omega_x P|, |1-F| \rangle^2. \tag{3.43}$$

Substituting the last term on the right hand side of (3.43) into (3.40a) yields (3.42), which is obtained if and only if equality holds in (3.43). In turn, equality in (3.43) is achieved iff $|\Omega_x A| = \kappa^2 |(1-F)PB|$, a.e. on $[-\pi, \pi]$, for arbitrary $\kappa^2 \in \mathbb{R}^+$. This equation, when combined with (3.40b) and (2.1) (see page 33), leads directly to (3.41). □

If $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$ satisfies condition i) in Assumption 3.1, then there exist stable filters $A(z)$ and $B(z)$ with the frequency response magnitudes given by (3.41). However, depending on $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$, the optimal frequency response magnitudes characterized (3.41) may be non-realizable, and can, in some cases, lead to unstable filters. In these situations, it is always possible to obtain a performance arbitrarily close to the optimal one by using stable filters, as shown in the next proposition.

**Lemma 3.4.** *Denote the frequency response magnitudes characterized by* (3.41a) *and* (3.41b) *by* $A_{inf}(\mathrm{e}^{j\omega})$ *and* $B_{inf}(\mathrm{e}^{j\omega})$*, respectively. If* $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$ *satisfies condition i) in Assumption 3.1, then* $A_{inf}(z)$ *and* $B_{inf}(z)$ *can be chosen stable; else, if* $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$ *satisfies condition ii) in Assumption 3.1, then one can achieve an FWMSE arbitrarily close to* $\sigma^2_{\epsilon\, inf|F}$ *with causal and stable filters* $A(z)$, $B(z)$ *such that*

$$\left|A(\mathrm{e}^{j\omega})\right| = A^{[\varepsilon]}(\omega) \triangleq \begin{cases} \varepsilon_B & ,\forall \omega \in \mathcal{I}_{\varepsilon_B} \\ 1/\varepsilon_A & ,\forall \omega \in \mathcal{I}_{\varepsilon_A} \\ \left|A_{inf}(\mathrm{e}^{j\omega})\right| & ,\forall \omega \notin \mathcal{I}_{\varepsilon_A} \cup \mathcal{I}_{\varepsilon_B}, \end{cases} \tag{3.44a}$$

$$\left|B(\mathrm{e}^{j\omega})\right| = B^{[\varepsilon]}(\omega) \triangleq \left(A^{[\varepsilon]}(\omega)\right)^{-1}, \tag{3.44b}$$

*a.e. on* $[-\pi, \pi]$*, where*

$$\mathcal{I}_{\varepsilon_B} \triangleq \{\omega \in [-\pi, \pi] : \left|B_{inf}(\mathrm{e}^{j\omega})\right| > \tfrac{1}{\varepsilon_B}\} \cup \mathcal{N}_P,$$
$$\mathcal{I}_{\varepsilon_A} \triangleq \{\omega \in [-\pi, \pi] : \left|A_{inf}(\mathrm{e}^{j\omega})\right| > \tfrac{1}{\varepsilon_A}\} \cup \mathcal{N}_{\Omega_{\mathrm{x}}},$$

*by making* $\varepsilon_A, \varepsilon_B \to 0$. ▲

*Proof.* We note that for any $\varepsilon_A, \varepsilon_B > 0$, the functions $A^{[\varepsilon]}, B^{[\varepsilon]} \in L^2$, and $A^{[\varepsilon]}(\omega), B^{[\varepsilon]}(\omega) > 0,\ \forall \omega \in [-\pi, \pi]$. As a consequence, one can always find causal, rational and stable filters $A(z)$ and $B(z)$ satisfying (3.44). Secondly, the difference between $\sigma^2_{\epsilon\, inf|F}$ and $\sigma^2_\epsilon$ when $\left|A(\mathrm{e}^{j\omega})\right|$ and $\left|B(\mathrm{e}^{j\omega})\right|$ satisfy (3.44) is given by

$$\sigma^2_\epsilon - \sigma^2_{\epsilon\, inf|F} = \frac{N^{[\varepsilon]} - N_{inf}}{\gamma - \|F\|^2}, \tag{3.45}$$

where $N^{[\varepsilon]} \triangleq \|\Omega_{\mathrm{x}} A^{[\varepsilon]}\|^2 \|P(1-F)B^{[\varepsilon]}\|^2$ and $N_{inf} \triangleq \|\Omega_{\mathrm{x}} A_{inf}\|^2 \|P(1-F)B_{inf}\|^2$. Defining

$$\begin{aligned} e_A(\mathrm{e}^{j\omega}) &\triangleq A^{[\varepsilon]}(\omega) - \left|A_{inf}(\mathrm{e}^{j\omega})\right|, \\ e_B(\mathrm{e}^{j\omega}) &\triangleq B^{[\varepsilon]}(\omega) - \left|B_{inf}(\mathrm{e}^{j\omega})\right|, \end{aligned}$$

and with $f(\mathrm{e}^{j\omega})$ as in (3.22), we can write

$$\begin{aligned} N^{[\varepsilon]} - N_{inf} &= \|\Omega_{\mathrm{x}}(|A_{inf}| + e_A)\|^2 \left\|fP\left(|B_{inf}| + e_B\right)\right\|^2 - \|\Omega_{\mathrm{x}} A_{inf}\|^2 \|fPB_{inf}\|^2 \\ &= \|\Omega_{\mathrm{x}} A_{inf}\|^2 \left(\|fPe_B\|^2 + 2\langle |P|^2 f^2 \left|B_{inf}\right|, e_B\rangle\right) \\ &\quad + \|fPB_{inf}\|^2 \left(\|\Omega_{\mathrm{x}} e_A\|^2 + 2\langle |\Omega_{\mathrm{x}}|^2 \left|A_{inf}\right|, e_A\rangle\right) \\ &= N_{inf}^{\frac{1}{2}} \Big[\|fP\, e_B\|^2 + \|\Omega_{\mathrm{x}} e_A\|^2 + 2\langle |P|^2 f^2 \left|B_{inf}\right|, e_B\rangle + 2\langle |\Omega_{\mathrm{x}}|^2 \left|A_{inf}\right|, e_A\rangle\Big]. \end{aligned}$$

Each of the terms above can be upper bounded as follows:

$$\begin{aligned} \|fP\, e_B\|^2 &\overset{(a)}{\leq} \int_{\mathcal{I}_{\varepsilon_A}} \left|P(\mathrm{e}^{j\omega})\right|^2 f(\mathrm{e}^{j\omega})^2 \varepsilon_A^2 d\omega + \int_{\mathcal{I}_{\varepsilon_B}} \left|P(\mathrm{e}^{j\omega})\right|^2 f(\mathrm{e}^{j\omega})^2 \left|B_{inf}(\mathrm{e}^{j\omega})\right|^2 d\omega \\ &\overset{(b)}{\leq} \varepsilon_A^2 \|fP\|^2 + \int_{\mathcal{I}_{\varepsilon_B}} \left|P(\mathrm{e}^{j\omega})\right| \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right| f(\mathrm{e}^{j\omega}) d\omega \\ &\overset{(c)}{\leq} \varepsilon_A^2 \|fP\|^2 + \varepsilon_B^2 \|\Omega_{\mathrm{x}}\|^2/\kappa^2, \\ \|\Omega_{\mathrm{x}} e_A\|^2 &\overset{(d)}{\leq} \int_{\mathcal{I}_{\varepsilon_B}} \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^2 \varepsilon_B^2 d\omega + \int_{\mathcal{I}_{\varepsilon_A}} \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^2 \left|A_{inf}(\mathrm{e}^{j\omega})\right|^2 d\omega \\ &\overset{(e)}{\leq} \varepsilon_B^2 \|\Omega_{\mathrm{x}}\|^2 + \int_{\mathcal{I}_{\varepsilon_A}} \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right| \left|P(\mathrm{e}^{j\omega})\right| f(\mathrm{e}^{j\omega}) d\omega \\ &\overset{(f)}{\leq} \varepsilon_B^2 \|\Omega_{\mathrm{x}}\|^2 + \varepsilon_A^2 \|fP\|^2 \kappa^2, \\ \langle |P|^2 f(\mathrm{e}^{j\omega})^2 \left|B_{inf}\right|, e_B\rangle &\overset{(g)}{\leq} \int_{\mathcal{I}_{\varepsilon_A}} \left|P(\mathrm{e}^{j\omega})\right|^2 f(\mathrm{e}^{j\omega})^2 \left|B_{inf}(\mathrm{e}^{j\omega})\right| \varepsilon_A d\omega \\ &\overset{(h)}{\leq} \varepsilon_A^2 \|fP\|^2, \\ \langle |\Omega_{\mathrm{x}}|^2 \left|A_{inf}\right|, e_A\rangle &\overset{(i)}{\leq} \int_{\mathcal{I}_{\varepsilon_B}} \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^2 \left|A_{inf}(\mathrm{e}^{j\omega})\right| \varepsilon_B d\omega \\ &\overset{(j)}{\leq} \varepsilon_B^2 \|\Omega_{\mathrm{x}}\|^2. \end{aligned}$$

In the above, $(a)$ follows from the fact that

$$\left|e_B(\mathrm{e}^{j\omega})\right| \leq \varepsilon_A, \forall \omega \in \mathcal{I}_{\varepsilon_A}, \qquad \text{and} \tag{3.46a}$$

$$-\left|B_{inf}(\mathrm{e}^{j\omega})\right| < e_B(\mathrm{e}^{j\omega}) < 0, \forall \omega \in \mathcal{I}_{\varepsilon_B}. \tag{3.46b}$$

$(b)$ follows from the fact that

$$\begin{aligned}\left|P(\mathrm{e}^{j\omega})\right|^2 \left|1 - F(\mathrm{e}^{j\omega})\right|^2 \left|B_{inf}(\mathrm{e}^{j\omega})\right|^2 \\ = \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^2 \left|A_{inf}(\mathrm{e}^{j\omega})\right|^2 \\ = \left|P(\mathrm{e}^{j\omega})\right| \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right| \left|1 - F(\mathrm{e}^{j\omega})\right|,\end{aligned} \tag{3.47}$$

$\forall \omega \in [-\pi, \pi]$, see (3.41), and from $\mathcal{I}_{\varepsilon_A} \subset [-\pi, \pi]$. Inequality $(c)$ follows from the fact that

$$\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right| < \varepsilon_A^2 \kappa^2 \left|P(\mathrm{e}^{j\omega})\right| f(\mathrm{e}^{j\omega}), \quad \forall \omega \in \mathcal{I}_{\varepsilon_A}; \tag{3.48a}$$

$$\left|P(\mathrm{e}^{j\omega})\right| < \varepsilon_B^2 \kappa^{-2} \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right| f(\mathrm{e}^{j\omega})^{\sim 1}, \quad \forall \omega \in \mathcal{I}_{\varepsilon_B}, \tag{3.48b}$$

which is readily obtained from (3.41) and (3.44). Inequality $(d)$ follows from

$$\left|e_A(\mathrm{e}^{j\omega})\right| \leq \varepsilon_B, \forall \omega \in \mathcal{I}_{\varepsilon_B}, \qquad \text{and} \tag{3.49a}$$

$$-\left|A_{inf}(\mathrm{e}^{j\omega})\right| < e_A(\mathrm{e}^{j\omega}) < 0, \forall \omega \in \mathcal{I}_{\varepsilon_A}. \tag{3.49b}$$

Inequality $(e)$ is due to (3.47) and to the fact that $\mathcal{I}_{\varepsilon_B} \subset [-\pi, \pi]$. Inequality $(f)$ stems from (3.48). Inequality $(g)$ follows from (3.46), while $(h)$ follows from the fact that $\left|B_{inf}(\mathrm{e}^{j\omega})\right| \leq \varepsilon_A$, $\forall \omega \in \mathcal{I}_{\varepsilon_A}$. Inequality $(i)$ stems from (3.49), while $(j)$ follows from the fact that $\left|A_{inf}(\mathrm{e}^{j\omega})\right| \leq \varepsilon_B$, $\forall \omega \in \mathcal{I}_{\varepsilon_B}$. Therefore,

$$\begin{aligned}N^{[\varepsilon]} - N_{inf} \\ \leq N_{inf}^{1/2} \left[(3 + \kappa^2)\|fP\|^2 \varepsilon_A^2 + (3 + \kappa^{-2})\|\Omega_{\mathrm{x}}\|^2 \varepsilon_B^2\right],\end{aligned}$$

which completes the proof. □

## 3.5 $F(z)$ Given

In this section we solve the problem of finding the filters $A(z)$ and $B(z)$ that minimize $D_{a,b}(\mathrm{x}, \mathrm{y})$ when the feedback filter $F(z)$ is given. More precisely, we seek the solution to the following

**Optimization Problem 3.4.** *For a given $K > 1$, and frequency response magnitudes $|\Omega_{\mathrm{x}}|$, $|P|$ and $f$, find the filters $A(z)$ and $B(z)$ that minimize*

$$D_{a,b}(\mathrm{x}, \mathrm{y}) = a \frac{\|\Omega_{\mathrm{x}} A\|^2 \|PBf\|^2}{K - \|f\|^2} + b\|(AB - 1)\Omega_{\mathrm{x}} P\|^2. \tag{3.50}$$

▲

**Theorem 3.5.** *For any given and fixed $F(z)$, the solution to Optimization Problem 3.4 is*

$$\left|B(\mathrm{e}^{j\omega})\right| = \frac{\left|P(\mathrm{e}^{j\omega})\right|^{\sim 1}}{\kappa}\sqrt{\frac{G(\mathrm{e}^{j\omega})}{f(\mathrm{e}^{j\omega})} - \xi}\,, \quad \textit{a.e. on } [-\pi,\pi] \tag{3.51a}$$

$$\left|A(\mathrm{e}^{j\omega})\right| = \kappa\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^{\sim 1}\sqrt{G(\mathrm{e}^{j\omega})f(\mathrm{e}^{j\omega}) - \xi f(\mathrm{e}^{j\omega})^2}, \quad \textit{a.e. on } [-\pi,\pi], \tag{3.51b}$$

$$\mathcal{R}\left\{A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right\} \geq 0, \qquad \mathcal{I}\left\{A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right\} = 0, \quad \textit{a.e. on } [-\pi,\pi] \tag{3.51c}$$

*where $\kappa > 0$ is an arbitrary real constant, $\xi$ is the unique scalar that satisfies*

$$K = \frac{a}{b}\frac{1}{2\pi}\int_{-\pi}^{\pi}\max\left\{f(\mathrm{e}^{j\omega})^2\,,\ \frac{f(\mathrm{e}^{j\omega})\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|}{\xi}\right\}d\omega + [1-\tfrac{a}{b}]\|f\|^2, \tag{3.52}$$

*and*

$$G(\mathrm{e}^{j\omega}) \triangleq \max\left\{\xi f(\mathrm{e}^{j\omega})\,,\ \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|\right\}, \quad \forall\omega\in[-\pi,\pi]. \tag{3.53}$$

*The scalar $\xi$ is related to $\sigma_{\mathrm{n}}^2$ and $\kappa$ via*

$$\xi = \frac{(a/b)\sigma_n^2}{\kappa^2} = \frac{(a/b)\|\Omega_{\mathrm{x}}A\|^2}{\kappa^2\left(K-\|f\|^2\right)}. \tag{3.54}$$

*The minimum of $D_{a,b}$ under the conditions of Optimization Problem 3.4, achieved if and only if $A(z)$ and $B(z)$ satisfy* (3.51)*, is*

$$\min_{A,B} D_{a,b}(\mathrm{x},\mathrm{y}) = \frac{b\xi}{2\pi}\int\limits_{|\Omega_{\mathrm{x}}P|\geq\xi f} f(\mathrm{e}^{j\omega})\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|d\omega + \frac{b}{2\pi}\int\limits_{|\Omega_{\mathrm{x}}P|<\xi f}\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2 d\omega \tag{3.55}$$

▲

*Proof.* From the proof of Theorem 3.1, the optimal signal transfer function $W(z)$ for a given $A(z)$ and $f(\mathrm{e}^{j\omega})$ satisfies (3.32). On the other hand, the optimal $A(z)$ for fixed $f$ and $W(z)$ is given by (3.41a). Since the optimal $A(z)$ and $W(z)$ must be reciprocally optimal, these transfer functions must satisfy (3.41a) and (3.32) simultaneously. From this fact, and substituting (3.41a) into (3.32), it follows that the optimal $W(z)$ satisfies

$$W(\mathrm{e}^{j\omega}) = \frac{\kappa^2 b\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\left|P(\mathrm{e}^{j\omega})\right| f(\mathrm{e}^{j\omega})\left|W(\mathrm{e}^{j\omega})\right|}{a\sigma_n^2 f(\mathrm{e}^{j\omega})^2 + \kappa^2 b\Omega_{\mathrm{x}}\left|P(\mathrm{e}^{j\omega})\right| f(\mathrm{e}^{j\omega})\left|W(\mathrm{e}^{j\omega})\right|}, \quad \text{a.e. on } [-\pi,\pi].$$

At any frequency $\omega\in[-\pi,\pi]$ (except possibly on a zero-measure subset of $[-\pi,\pi]$), the frequency response $W(\mathrm{e}^{j\omega})$ that satisfies this equation must satisfy either

$$W(\mathrm{e}^{j\omega}) = 0, \tag{3.56}$$

or else

$$W(\mathrm{e}^{j\omega}) = \max\left\{0, 1 - \xi' \frac{f(\mathrm{e}^{j\omega})}{|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})|}\right\}, \tag{3.57}$$

where

$$\xi' \triangleq \frac{(a/b)\sigma_n^2}{\kappa^2} = \frac{(a/b)\|\Omega_{\mathrm{x}}A\|^2}{\kappa^2\,(K - \|f\|^2)} = \frac{(a/b)\langle|\Omega_{\mathrm{x}}P|\,f, |W|\rangle}{K - \|f\|^2}. \tag{3.58}$$

The last two equalities in (3.58) follow by substituting (3.20) and (3.41a) into the left hand side of (3.58). Next we show that the optimal $W(z)$ satisfies (3.57) (and *not* (3.56)) almost everywhere on $[-\pi, \pi]$. For this purpose, we write (3.42) as

$$D_{a,b}(\mathrm{x}, \mathrm{y}) = \mathscr{W}(W) \triangleq a\frac{\langle|\Omega_{\mathrm{x}}P|\,f, W\rangle^2}{K - \|f\|^2} + b\|(W-1)\Omega_{\mathrm{x}}P\|^2 \tag{3.59}$$

If $W(\mathrm{e}^{j\omega})$ does not satisfy (3.57) almost everywhere on $[-\pi, \pi]$, then there exists a non-zero measure set of frequencies $\mathbb{W}$ such that $W(\mathrm{e}^{j\omega}) = 0$ and $\xi' f(\mathrm{e}^{j\omega}) < \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$, for all $\omega \in \mathbb{W}$. To show that such an $W(z)$ is not optimal, we will demonstrate that the Gateaux differential of $\mathscr{W}(W)$, that is, $\delta\mathscr{W}(W; h)$, is negative for some choices of the function $h$. Applying (2.2) (page 33) to (3.59) we find that

$$\delta\mathscr{W}(W; h) = 2a\frac{\langle|\Omega_{\mathrm{x}}P|\,f, W\rangle}{K - \|f\|^2}\langle|\Omega_{\mathrm{x}}P|\,f, h\rangle + 2b\langle(W-1)\,|\Omega_{\mathrm{x}}P|^2, h\rangle.$$

Let us choose $h$ to be such that $h(\omega) = 0$, $\forall\omega \notin \mathbb{W}$ and such that $h(\omega) > 0$, $\forall\omega \in \mathbb{W}$. Then

$$\delta\mathscr{W}(W; h) = 2b\left(\langle\xi' f\,|\Omega_{\mathrm{x}}P|, h(\omega)\rangle - \langle|\Omega_{\mathrm{x}}P|^2, h\rangle\right) = 2b\langle[\xi' f - |\Omega_{\mathrm{x}}P|\,, |\Omega_{\mathrm{x}}P|\,h]\rangle < 0.$$

where the inequality follows from our initial supposition on $\mathbb{W}$, which implies that $\xi' f(\mathrm{e}^{j\omega}) < \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$ over a non-zero measure set of frequencies. Thus, the optimal $W(z)$ is such that $W(\mathrm{e}^{j\omega})$ satisfies (3.57) a.e. on $[-\pi, \pi]$.

In order to obtain an explicit solution for the optimal $W(z)$, we need to express $\xi'$ only in terms of $\Omega_{\mathrm{x}}$, $P$ and $f$. For this purpose, define

$$G'(\mathrm{e}^{j\omega}) \triangleq \max\left\{\xi' f(\mathrm{e}^{j\omega})\,,\ \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|\right\}, \quad \forall\omega \in [-\pi, \pi] \tag{3.60}$$

where $\xi'$ is as in (3.58). With this definition, (3.57) can be written as

$$W(\mathrm{e}^{j\omega}) = 1 - \xi' f(\mathrm{e}^{j\omega})/G'(\omega). \tag{3.61}$$

Substituting the latter into the right hand side of (3.58) we obtain

$$\xi' = (a/b)\frac{\langle|\Omega_{\mathrm{x}}P|\,f, 1 - \xi' f G'^{-1}\rangle}{K - \|f\|^2} = (a/b)\frac{\langle G' f, 1 - \xi' f G'^{-1}\rangle}{K - \|f\|^2} = (a/b)\frac{\langle G', f\rangle - \xi'\|f\|^2}{K - \|f\|^2}. \tag{3.62}$$

Solving for $\xi'$ in (3.62) it is found that

$$\xi' = \frac{(a/b)\langle G', f\rangle}{K - \left[1 - \frac{a}{b}\right]\|f\|^2}. \tag{3.63}$$

It is easy to verify that there exists a unique value for the scalar $\xi'$ that satisfies (3.60) and (3.63) simultaneously. Noting that (3.60) and (3.63) are equivalent to (3.52) and (3.53), respectively, this implies that

$$\xi' = \xi, \qquad G'(\mathrm{e}^{j\omega}) = G(\mathrm{e}^{j\omega}), \quad \forall\omega \in [-\pi, \pi]. \tag{3.64}$$

Substitution of (3.64) into (3.63) and then into (3.61) yields

$$W(\mathrm{e}^{j\omega}) = 1 - \frac{(a/b)\langle G, f\rangle}{K - \left[1 - \frac{a}{b}\right]\|f\|^2} \cdot \frac{f(\mathrm{e}^{j\omega})}{G(\mathrm{e}^{j\omega})}. \tag{3.65}$$

Substitution of (3.65) into (3.41a) and (3.41b) yields (3.51c) and (3.51a), respectively. Substitution of (3.65) into the right hand side of (3.42) yields

$$\begin{aligned}
&\min_{A,B} D_{a,b}(\mathrm{x}, \mathrm{y}) \\
&= a\frac{\left\langle f\,, |\Omega_{\mathrm{x}}P|\left(1 - \frac{(a/b)\langle G,f\rangle}{K-[1-\frac{a}{b}]\|f\|^2}\frac{f}{G}\right)\right\rangle^2}{K - \|f\|^2} + b\left\|\frac{(a/b)\langle G,f\rangle}{K-[1-\frac{a}{b}]\|f\|^2}\cdot\frac{f}{G}\Omega_{\mathrm{x}}P\right\|^2 \\
&= a\frac{\left[\langle G,f\rangle - \frac{(a/b)\langle G,f\rangle}{K-[1-\frac{a}{b}]\|f\|^2}\|f\|^2\right]^2}{K-\|f\|^2} + \frac{a^2}{b}\frac{\langle G,f\rangle^2\left(\|f\|^2 - \|(1-I_f)f\|^2\right)}{\left(K-[1-\frac{a}{b}]\|f\|^2\right)^2} + b\|(1-I_f)\Omega_{\mathrm{x}}P\|^2 \\
&= a\langle G,f\rangle^2\left(\frac{\left(K-[1-\frac{a}{b}]\|f\|^2 - \frac{a}{b}\|f\|^2\right)^2}{(K-\|f\|^2)\left(K-[1-\frac{a}{b}]\|f\|^2\right)^2} + \frac{\frac{a}{b}\|f\|^2}{\left(K-[1-\frac{a}{b}]\|f\|^2\right)^2}\right) \\
&\qquad\qquad - b\xi^2\|(1-I_f)f\|^2 + b\|(1-I_f)\Omega_{\mathrm{x}}P\|^2 \\
&= a\frac{\langle G,f\rangle^2}{\left(K-[1-\frac{a}{b}]\|f\|^2\right)^2}\left(K - \|f\|^2 + \tfrac{a}{b}\|I_f f\|^2\right) - b\xi^2\|(1-I_f)f\|^2 + b\|(1-I_f)\Omega_{\mathrm{x}}P\|^2 \\
&= a\frac{\langle G,f\rangle^2}{K-[1-\frac{a}{b}]\|f\|^2} - b\xi^2\|(1-I_f)f\|^2 + b\|(1-I_f)\Omega_{\mathrm{x}}P\|^2,
\end{aligned} \tag{3.66}$$

where the indicator function, $I_f(\mathrm{e}^{j\omega})$, is defined as

$$I_f(\mathrm{e}^{j\omega}) \triangleq \begin{cases} 1 & , \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| \geq \xi f(\mathrm{e}^{j\omega}) \\ 0 & , \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| < \xi f(\mathrm{e}^{j\omega}). \end{cases}$$

Rearranging terms in (3.66) leads directly to (3.55). From (3.64) and (3.60), it follows that (3.63) is equivalent to

$$K = \frac{a}{b}\frac{\langle G, f\rangle}{\xi} + \left[1 - \frac{a}{b}\right]\|f\|^2. \tag{3.67}$$

Substitution of (3.53) into (3.67) yields (3.52). Also, from (3.64) and (3.58), we have that

$$\xi = \frac{(a/b)\sigma_n^2}{\kappa^2} = \frac{(a/b)\|\Omega_{\mathrm{x}} A\|^2}{\kappa^2\,(K - \|f\|^2)}$$

This completes the proof. □

When $a = b = 1$, i.e., when WCMSE equals MSE, the optimal filters characterized by Theorem 3.5 correspond to those found in [87]. The result in [87] was obtained solving an *iso-perimetric* problem [139, 140] by means of Lagrange multipliers. Interestingly, the proof of Theorem 3.5, apart from using a simple variational argument, is based only upon algebraic manipulation.

In addition, choosing $a = b = 1$, and assuming $f(\mathrm{e}^{j\omega})$ is such that $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| \geq \xi f(\mathrm{e}^{j\omega})$, $\forall \omega \in [-\pi, \pi]$, we have that (3.55) becomes

$$D_{1,1}(\mathrm{x}, \mathrm{y}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\mathrm{e}^{j\omega})\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| d\omega \tag{3.68}$$

Since $f(\mathrm{e}^{j\omega})$ needs to satisfy (3.76), it is straightforward to show that the function $f(\mathrm{e}^{j\omega})$ that minimizes (3.68) satisfies

$$f(\mathrm{e}^{j\omega}) = \frac{\eta^2_{\Omega_{\mathrm{x}}P}}{|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})|}, \tag{3.69}$$

where

$$\eta^2_{\Omega_{\mathrm{x}}P} \triangleq \mathrm{e}^{\frac{1}{\pi}\int_{-\pi}^{\pi} \ln\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| d\omega} \tag{3.70}$$

is the *minimal prediction variance* of a w.s.s. process having PSD $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2$. With this result, the condition $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| \geq \xi f(\mathrm{e}^{j\omega})$, $\forall \omega \in [-\pi, \pi]$ becomes

$$\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2 \geq \frac{\eta^2_{\Omega_{\mathrm{x}}P}}{K} \tag{3.71}$$

The optimal filters $A(z)$ and $B(z)$ are then characterized by substituting (3.69) into (3.51). The result is the same as the one derived by the author and colleagues in [141]. This result also happens to be a special case of the filters first characterized by Zamir, Kochman and Erez first in [142] and then in [15]. We shall generalize (3.69) beyond the assumption (3.71) and for general choices of $a$ and $b$ later in Section 3.9.1.

## 3.6 $A(z)$ and $B(z)$ Given

In this case, the minimization of the WCMSE reduces to the following:

**Optimization Problem 3.5.** *For a given $K > 1$ and given frequency responses $\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})$, $P(\mathrm{e}^{j\omega})$, $B(\mathrm{e}^{j\omega})$, $A(\mathrm{e}^{j\omega})$, find a frequency response magnitude $f(\mathrm{e}^{j\omega})$ so as to*

$$\textit{minimize:} \quad D_{a,b}(\mathrm{x},\mathrm{y}) = a\frac{\|\Omega_{\mathrm{x}}A\|^2\|PBf\|}{K-\|f\|^2} + b\|(AB-1)\Omega_{\mathrm{x}}P\|^2 \tag{3.72}$$

$$\textit{subject to:} \quad \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(\mathrm{e}^{j\omega})d\omega \geq 0. \tag{3.73}$$

$$\|f\|^2 < K,$$

$$f(\mathrm{e}^{j\omega}) \geq 0, \quad \forall\omega \in [-\pi,\pi]. \tag{3.74}$$

▲

Notice that the above optimization problem is stated in terms of $f(\mathrm{e}^{j\omega})$ (see (3.22) on page 53), and not directly in terms of $F(z)$. It is therefore necessary to guarantee that searching over all functions $f(\mathrm{e}^{j\omega})$ that could yield a solution is equivalent to search over all filters $F(z)$ that are feasible solutions. For this purpose, we next translate the restrictions on $F(z)$, stated in Constraints 3.1 and 3.2, into equivalent constraints on $f$. To begin with, note that, by definition, $f$ needs to satisfy

$$f(\mathrm{e}^{j\omega}) \geq 0, \forall\omega \in [-\pi,\pi], \tag{3.75}$$

and that, since $\|F\|^2 = \|F-1\|^2 - 1$, see (3.19), Constraint 3.2 is satisfied iff $\|f\|^2 < \gamma + 1$. In addition, a stable and strictly causal $F(z)$ (i.e., one satisfying Constraint 3.1) always leads to a function $f$, see (3.22), which satisfies[7]

$$0 \leq \int_{-\pi}^{\pi} \ln f(\mathrm{e}^{j\omega})d\omega < \infty. \tag{3.76}$$

This result follows directly from Jensen's formula [144] (see also the Bode Integral Theorem in, e.g., [145]).

On the other hand, as Theorem 3.7 will show, if Assumption 3.1 holds, then the optimal $f$ within the set of functions described by (3.76) and the requirement $\|f\|^2 < \gamma + 1$ turns out to be piece-wise differentiable on $[-\pi,\pi]$, has at most a finite number of discontinuity points, and satisfies

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\mathrm{e}^{j\omega})d\omega = 0, \quad \text{and} \tag{3.77a}$$

$$0 < f_{min} \leq f(\mathrm{e}^{j\omega}) \leq f_{max} < \infty, \forall\omega \in [-\pi,\pi]. \tag{3.77b}$$

Under these conditions, it is always possible to find a stable and strictly causal filter $F(z)$ such that $\left|1 - F(\mathrm{e}^{j\omega})\right|$ approximates $f(\mathrm{e}^{j\omega})$ arbitrarily well on $[-\pi,\pi]$, as stated in the following lemma:

[7] Notice that (3.76) dictates a fundamental trade-off in the noise-shaping capabilities of feedback quantizers, namely, that one can remove noise from one frequency band only at the expense of increasing it in another. This is also known as the "water-bed effect", see, e.g., [143]. We will discuss further implications of (3.76) in Section 3.9.2.

**Lemma 3.6.** *Suppose that $f$ is piece-wise differentiable on $[-\pi,\pi]$, that it has at most a finite number of discontinuity points and that it satisfies* (3.77)*. Then, for every $\varepsilon > 0$, there exists a (finite order) rational, strictly proper and stable $F(z)$ such that $\|f - |1-F|\| \leq \varepsilon$.* ▲

*Proof.* Define the partition $-\pi = \omega_0 < \omega_1 < \cdots < \omega_p = \pi$, where $\{\omega_i\}_{i=1}^{p-1}$ correspond to the discontinuity points (if any) of $f$. Since $f$ is piece-wise differentiable, its first derivative over all open intervals $(\omega_i, \omega_{i+1})$, $i \in \{0,\ldots,p-1\}$ is bounded by a constant $0 \leq S < \infty$. For each $m > S$, we define the set $\mathcal{T}_m$, consisting of all *continuous* functions $h : [-\pi,\pi] \to \mathbb{R}^+$ satisfying

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log h(\omega) d\omega = 0, \tag{3.78a}$$

$$f_{min} \leq h(\omega) \leq f_{max}, \quad \forall \omega \in [-\pi,\pi], \text{ and} \tag{3.78b}$$

$$\left|\frac{d}{d\omega} h(\omega)\right| \leq m, \ \forall \omega \in [-\pi,\pi]. \tag{3.78c}$$

For each $m$, the function

$$h_m \triangleq \arg\min_{h\in\mathcal{T}_m} \|f-h\|. \tag{3.79}$$

is the element in $\mathcal{T}_m$ "closest" to $f$. From (3.77b), and from the fact that $f$ is piece-wise differentiable, it follows that for every $\varepsilon_0 > 0$, there exists a bounded $T \geq S$ such that

$$\|f - h_m\| \leq \varepsilon_0, \quad \forall m > T. \tag{3.80}$$

(Indeed, it is easy to obtain the bound $\|f(\mathrm{e}^{j\omega}) - h_m(\omega)\| \leq (f_{max} - f_{min})^2 p/m$). Notice that if $f$ had no discontinuity points and if $m \geq S$, then $h_m \equiv f$ (see (3.78c)), yielding $\|f - h_m\| = 0$.

Since $h_m(\omega)$ is continuous and piece-wise differentiable, its Fourier series converges uniformly over $[-\pi,\pi]$. Thus, by definition, for every $\varepsilon_1 > 0$, there exists an $N$-th order (where $N < \infty$ is odd and depends on $\varepsilon_1$) rational transfer function $H_N(z)$ (the $Z$-transform of the coefficients of the $\frac{N-1}{2}$-th partial sum of the Fourier series of $f$) such that

$$\left|h_m(\omega) - H_N(\mathrm{e}^{j\omega})\right| < \varepsilon_1, \ \forall \omega \in [-\pi,\pi]. \tag{3.81}$$

$H_N(z)$ can be written as $H_N(z) = G_1 z^{-\frac{N+1}{2}} \prod_{i=1}^{N}(z-c_i)$, where $G_1 \in \mathbb{R}$. Thus, the transfer function

$$H'_N(z) \triangleq H_N(z) \frac{G_1}{|G_1|} z^{-\frac{N-1}{2}} \prod_{\substack{i=1,\\ |c_i|>1}}^{N} \frac{c_i}{|c_i|}\left(\frac{c_i^* z - 1}{z - c_i}\right)$$

is clearly biproper [8], stable, minimum-phase and such that $\left|H'_N(\mathrm{e}^{j\omega})\right| = \left|H_N(\mathrm{e}^{j\omega})\right|$, $\forall\omega \in [-\pi,\pi]$, with the first value of its impulse response being

$$\chi \triangleq \lim_{z\to\infty} H'_N(z) > 0.$$

Define $\tilde{H}_N(z) \triangleq \frac{1}{\chi}H'_N(z)$, so that $\lim_{z\to\infty}\tilde{H}_N(z) = 1$ and

$$\left|\tilde{H}_N(\mathrm{e}^{j\omega})\right| = \frac{1}{\chi}\left|H_N(\mathrm{e}^{j\omega})\right|, \quad \forall\omega\in[-\pi,\pi]. \tag{3.82}$$

With the choice $F(z) = 1-\tilde{H}_N(z)$, we have

$$\begin{aligned}\|f-|1-F|\| = \left\|f-|\tilde{H}_N|\right\| &\le \|f-h_m\| + \left\|h_m-|\tilde{H}_N|\right\| \\ &\le \varepsilon_0+\varepsilon_1+\left\|h_m-|\tilde{H}_N|\right\|.\end{aligned} \tag{3.83}$$

We now proceed to find an upper bound for the last term in the above inequality. From (3.81) and (3.82), we have that

$$\begin{aligned}\left\|h_m-|\tilde{H}_N|\right\| &\le \|h_m-|H_N|\| + \left\||H_N|-|\tilde{H}_N|\right\| \\ &\le \varepsilon_1 + \left|1-\frac{1}{\chi}\right|\|H_N\| = \varepsilon_1+\frac{|\chi-1|}{\chi}\|H_N\|.\end{aligned} \tag{3.84}$$

From Jensen's formula (see, e.g., [144]), and since $H'_N(z)$ is stable and minimum phase, we obtain

$$\log\chi = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left|H'_N(\mathrm{e}^{j\omega})\right|d\omega. \tag{3.85}$$

Recalling from (3.78a) and (3.79) that $\frac{1}{2\pi}\int_{-\pi}^{\pi}\log h_m(\omega)d\omega = 0$, we can write (3.85) as

$$\log\chi = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left(\frac{\left|H_N(\mathrm{e}^{j\omega})\right|}{h_m(\omega)}\right)d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left(\frac{h_m(\omega)+e(\omega)}{h_m(\omega)}\right)d\omega, \tag{3.86}$$

where $e(\omega) \triangleq \left|H_N(\mathrm{e}^{j\omega})\right| - h_m(\omega)$. From (3.81), we have that

$$|e(\omega)| = \left|h_m(\omega)-\left|H_N(\mathrm{e}^{j\omega})\right|\right| \le \left|h_m(\omega)-H_N(\mathrm{e}^{j\omega})\right| \le \varepsilon_1.$$

Thus, choosing $\varepsilon_1 < f_{min}$, the last integral in (3.86) can be upper and lower bounded as

$$\log\left(\frac{f_{min}-\varepsilon_1}{f_{min}}\right) \le \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left(\frac{h_m(\omega)+e(\omega)}{h_m(\omega)}\right)d\omega \le \log\left(\frac{f_{min}+\varepsilon_1}{f_{min}}\right).$$

It then follows from (3.86) that

$$1-\frac{\varepsilon_1}{f_{min}} \le \chi \le 1+\frac{\varepsilon_1}{f_{min}} \iff |\chi-1| \le \frac{\varepsilon_1}{f_{min}}$$

[8] A transfer function $F(z)$ is said to be biproper if and only if $0 < |\lim_{z\to\infty}F(z)| < \infty$.

Substituting the latter into (3.84), we obtain

$$\left\| h_m - |\tilde{H}_N| \right\| \leq \varepsilon_1 + \frac{\varepsilon_1}{f_{min} - \varepsilon_1} \|H_N\| \leq \varepsilon_1 + \frac{\varepsilon_1}{f_{min} - \varepsilon_1} (\|f\| + \varepsilon_0 + \varepsilon_1), \tag{3.87}$$

where the last inequality stems from (3.80) and (3.81). Substitution of (3.87) into (3.83) yields

$$\|f - |1 - F|\| \leq \varepsilon_0 + \varepsilon_1 + \frac{\varepsilon_1}{f_{min} - \varepsilon_1} (\|f\| + \varepsilon_0 + \varepsilon_1). \tag{3.88}$$

Since $\|f\|$ is bounded, and from (3.77b), it follows from (3.88) that for any $\varepsilon > 0$, one can always choose sufficiently large (bounded) values for $T$ (see (3.80)) and $N$ (see (3.81)) so that $\varepsilon_0$ and $\varepsilon_1$ are small enough to yield $\|f - |1 - F|\| < \varepsilon$. This completes the proof. □

Having verified that a feasible filter $F(z)$ can always be found to match almost any frequency response $f(\mathrm{e}^{j\omega})$ to any degree of accuracy, we present the solution to Optimization problem 3.5 in the following theorem:

**Theorem 3.7.** *The function $f$ that solves Optimization Problem 3.5 is given by*

$$f(\mathrm{e}^{j\omega}) = \sqrt{\frac{K\lambda}{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 + \lambda}}, \quad \textit{a.e. on } [-\pi, \pi], \tag{3.89a}$$

*where the parameter $\lambda > 0$ is the unique scalar satisfying*

$$\ln(K) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \left( \frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{\lambda} + 1 \right) d\omega. \tag{3.89b}$$

*The corresponding minimum of $D_{a,b}$ is given by*

$$\min_{\substack{f:\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(\mathrm{e}^{j\omega})d\omega \geq 0 \\ f(\mathrm{e}^{j\omega}) \geq 0,\ \forall \omega \in [-\pi,\pi]}} D_{a,b}(\mathrm{x}, \mathrm{y}) = a\lambda \|\Omega_{\mathrm{x}} A\|^2 + b\|(AB - 1)\Omega_{\mathrm{x}} P\|^2. \tag{3.89c}$$

▲

*Proof.* We first define and solve a related (but simpler) optimization problem, namely: For any given constant $C \in (1, K)$ and transfer functions $P(z)$, $B(z)$,

$$\text{minimize: } \mathscr{J}(f) \triangleq \|PBf\|^2 \tag{3.90a}$$

$$\text{subject to: } 0 \geq \mathscr{G}_1(f) \triangleq \|f\|^2 - C \tag{3.90b}$$

$$0 \geq \mathscr{G}_2(f) \triangleq -\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln f(\mathrm{e}^{j\omega}) d\omega \tag{3.90c}$$

$$0 \leq f(\mathrm{e}^{j\omega}),\ \forall \omega \in [-\pi, \pi].$$

Clearly, the functionals $\mathscr{J}$, $\mathscr{G}_1$ and $\mathscr{G}_2$ are convex. Moreover, for all $C > 1$, there exists a function $f_1 \in L^2$ such that $\mathscr{G}_1(f_1) < 0$ and $\mathscr{G}_2(f_1) < 0$ (a trivial example is $f_1(\omega) \equiv \sqrt{(C+1)/2}$). This allows one to apply Theorem 3.17 (page 105). From the latter, we have that the minimizer of $\mathscr{J}(f)$ subject to (3.90b) and (3.90c) is an extremizer of the Lagrangian

$$\begin{aligned}\mathscr{L}(f) &\triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 f(\mathrm{e}^{j\omega})^2 d\omega + \lambda_1 \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\mathrm{e}^{j\omega})^2 d\omega - \lambda_2 \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(\mathrm{e}^{j\omega}) d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 + \lambda_1\right) f(\mathrm{e}^{j\omega})^2 - \lambda_2 \ln f(\mathrm{e}^{j\omega}) d\omega\end{aligned}$$

for some $\lambda_1, \lambda_2 \geq 0$. An extremizer of $\mathscr{L}(f)$ must be such that its Gateaux differential satisfies

$$\delta\mathscr{L}(f;h) = 0 \text{ for all } h \in L^2. \tag{3.91}$$

Applying the definition given in (2.2) (page 33) to $\mathscr{L}(f)$ we obtain

$$\begin{aligned}\delta\mathscr{L}(f;h) &= \frac{\partial\mathscr{L}(f+\alpha h)}{\partial\alpha}\Big|_{\alpha=0} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left[2\left(\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 + \lambda_1\right) f(\mathrm{e}^{j\omega}) - \lambda_2 f(\mathrm{e}^{j\omega})^{-1}\right] h(\omega) d\omega.\end{aligned} \tag{3.92}$$

Notice that $f(\mathrm{e}^{j\omega})$ must necessarily be strictly positive a.e. on $[-\pi, \pi]$, since otherwise constraints (3.90c) and (3.90b) would not be met. This guarantees that (3.92) is well defined. It is clear from (3.92) that (3.91) holds iff

$$\begin{aligned}&0 = 2\left(\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 + \lambda_1\right) f(\mathrm{e}^{j\omega}) - \lambda_2 f(\mathrm{e}^{j\omega})^{-1}, \text{ a.e. on } [-\pi,\pi] \\ \iff\ & f(\mathrm{e}^{j\omega})^2 = \frac{\lambda_2/2}{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 + \lambda_1}, \text{ a.e. on } [-\pi,\pi].\end{aligned} \tag{3.93}$$

Notice from the latter that, if $\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right| = 0$ over a non-zero measure set of frequencies, then $\lambda_1$ cannot equal zero. On the other hand, if $\lambda_2 = 0$, then constraint (3.90b) could not be met. In view of (3.217) in Theorem 3.17 (page 105), this fact implies that constraint (3.90c) is satisfied with equality, i.e., we have:

$$\ln(\lambda_2/2) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 + \lambda_1\right) d\omega. \tag{3.94}$$

Therefore, the minimizer of $\mathscr{J}(f)$, satisfies

$$f(\mathrm{e}^{j\omega})^2 = f_{\lambda_1}(\omega)^2 \triangleq \frac{\beta(\lambda_1)}{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right| + \lambda_1}, \quad \text{a.e. on } [-\pi,\pi], \tag{3.95}$$

for some $\lambda_1 \geq 0$, where

$$\beta(\lambda_1) \triangleq \mathrm{e}^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2+\lambda_1\right)d\omega}. \tag{3.96}$$

It is clear from (3.96) that if $|PB|$ is almost constant, then $f_{\lambda_1}(\omega) = 1$, a.e. on $[-\pi, \pi]$, for all $\lambda_1 \geq 0$. Conversely, if $|PB|$ is *not* almost constant, then the value of $\lambda_1$ for which $f_{\lambda_1}$ solves (3.90) has to be found. For this purpose, we substitute (3.95) into (3.90a), obtaining

$$\mathscr{J}(f_{\lambda_1}) \triangleq J(\lambda_1) = \beta(\lambda_1)\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{s(\lambda_1,\omega)}d\omega, \tag{3.97}$$

where

$$s(\lambda_1,\omega) \triangleq \left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 + \lambda_1, \quad \forall\omega \in [-\pi,\pi], \forall\lambda_1 \geq 0$$

Differentiation of $J(\lambda_1)$ with respect to $\lambda_1$ yields

$$\begin{aligned}
&\frac{dJ(\lambda_1)}{d\lambda_1} \\
&= \beta_1(\lambda_1)\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega\int_{-\pi}^{\pi}\frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{s(\lambda_1,\omega)}d\omega - \beta(\lambda_1)\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{s(\lambda_1,\omega)^2}d\omega \\
&= \beta_1(\lambda_1)\left[\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega\int_{-\pi}^{\pi}\frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{s(\lambda_1,\omega)}d\omega - \frac{1}{2\pi}\left(\int_{-\pi}^{\pi}\frac{d\omega}{s(\lambda_1,\omega)} - \int_{-\pi}^{\pi}\frac{\lambda_1}{s(\lambda_1,\omega)^2}d\omega\right)\right] \\
&= \beta_1(\lambda_1)\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{s(\lambda_1,\omega)}d\omega - 1\right) + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\lambda_1}{s(\lambda_1,\omega)^2}d\omega\right] \\
&= \lambda_1\beta_1(\lambda_1)\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)^2}d\omega - \left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega\right)^2\right] \geq 0, \quad \forall\lambda_1 \geq 0.
\end{aligned} \tag{3.98}$$

Since $dJ(\lambda_1)/d\lambda_1 < 0$, it is clear that the minimizer of $\mathscr{J}(f)$ is given by (3.95) with $\lambda_1$ taking the smallest value allowed by the constraint

$$C \geq \beta(\lambda_1)\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega.$$

This arises by substituting (3.95) into (3.90b). Let the dependence of $\|f_{\lambda_1}\|^2$ on $\lambda_1$ be made explicit by the function

$$c(\lambda_1) \triangleq \|f_{\lambda_1}\|^2 = \beta(\lambda_1)\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega. \tag{3.99}$$

We have that

$$\frac{dc(\lambda_1)}{d\lambda_1} = \beta(\lambda_1)\left[\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega\right)^2 - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)^2}d\omega\right] \leq 0, \tag{3.100}$$

where the above inequality follows from Theorem 3.18 (in page 106) and the fact that $\beta(\lambda_1) > 0$, $\forall \lambda_1 \geq 0$. The inequality is strict if and only if $|PB|$ is not almost constant (which is the case). Thus, the function $c(\lambda_1)$ is monotonically decreasing with $\lambda_1$. Combining this with (3.98), it follows that, in order for $f_{\lambda_1}$ to be the solution of (3.90), the Lagrange multiplier $\lambda_1$ must be

$$\lambda_1 = \begin{cases} 0 & , \text{ if } C \geq c(0), \\ c^{-1}(C) & , \text{ if } C \leq c(0), \end{cases} \tag{3.101}$$

where $c^{-1}(\cdot)$ is the inverse of the function $c(\cdot)$ defined in (3.99). Also note from (3.100), (3.101) and the fact that $|PB|$ is *not* almost constant that, if $C < c(0)$, then $c^{-1}(C) > 0$, and thus the optimal $\lambda_1$ is strictly positive. Therefore, the (squared) function $f$ that solves (3.90) is given by (3.95), where $\lambda_1 \geq 0$ is the unique scalar that satisfies (3.101). Furthermore, we conclude from (3.99) and (3.101) that

$$\text{if } f^o = \arg \min_{f:\mathscr{G}_1(f),\mathscr{G}_2(f)\leq 0} \mathscr{J}(f), \qquad \text{then} \qquad \|f^o\|^2 \leq c(0). \tag{3.102}$$

Next, the solution to (3.90), given by (3.95) and (3.101), will be used to solve Optimization Problem 3.5. For this purpose, define the functional

$$\mathscr{V}(f) \triangleq \frac{\|PBf\|^2}{K - \|f\|^2}, \tag{3.103}$$

and let $f^\star$ be the minimizer of $\mathscr{V}(f)$ subject to (3.73) and (3.74). Define $C^\star \triangleq \|f^\star\|^2$. Then

$$\mathscr{V}(f^\star) = \min_{\substack{f:\, 0\geq \mathscr{G}_2(f) \\ 0\leq f(\mathrm{e}^{j\omega}),\ \forall \omega\in[-\pi,\pi] \\ \|f\|^2 = C^\star}} \frac{\mathscr{J}(f)}{K - C^\star} \geq \min_{\substack{f:\, 0\geq \mathscr{G}_2(f) \\ 0\leq f(\mathrm{e}^{j\omega}),\ \forall \omega\in[-\pi,\pi] \\ \|f\|^2 \leq C^\star}} \frac{\mathscr{J}(f)}{K - \|f\|^2} \tag{3.104}$$

In view of (3.102) the inequality in (3.104) is strict unless $C^\star \leq c(0)$. Moreover

$$f^\star = f_{\lambda_1^\star},$$

where $\lambda_1^\star$ is such that $c(\lambda_1^\star) = C^\star$. To find $\lambda_1^\star$, we substitute $f_{\lambda_1}$ into (3.103), which yields

$$\mathscr{V}(f_{\lambda_1}) = \Phi(\lambda_1) \triangleq \frac{\beta_1(\lambda_1)\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|PB|^2}{s(\lambda_1,\omega)} d\omega}{K - c(\lambda_1)}. \tag{3.105}$$

Notice that Optimization Problem 3.5 has been reduced to finding the scalar $\lambda_1^\star$ that minimizes the func-

tion $\Phi(\lambda_1)$. For this purpose, we simply differentiate $\Phi(\lambda_1)$ and make use of (3.90a) to obtain

$$\begin{aligned}
\frac{d\Phi(\lambda_1)}{d\lambda_1} &= \left[ c(\lambda_1)\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|PB|^2}{s(\lambda_1,\omega)}d\omega - \beta(\lambda_1)\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|PB|^2}{s(\lambda_1,\omega)^2}d\omega \right](K-c(\lambda_1)) \\
&\quad + \left[ \beta(\lambda_1)\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega\right)^2 - \beta(\lambda_1)\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)^2}d\omega \right] J(\lambda_1) \\
&= \left[ \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega\int_{-\pi}^{\pi}\frac{|PB|^2}{s(\lambda_1,\omega)}d\omega - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|PB|^2}{s(\lambda_1,\omega)^2}d\omega \right]\beta(\lambda_1)(K-c(\lambda_1)) \\
&\quad + \left[ \left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega\right)^2 - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)^2}d\omega \right]\beta(\lambda_1)J(\lambda_1) \\
&= \frac{\beta(\lambda_1)}{4\pi^2}\int_{-\pi}^{\pi}\frac{1}{s(\lambda_1,\omega)}d\omega\int_{-\pi}^{\pi}\frac{|PB|^2(K-c(\lambda_1))+J(\lambda_1)}{s(\lambda_1,\omega)}d\omega \\
&\qquad - \frac{\beta(\lambda_1)}{2\pi}\int_{-\pi}^{\pi}\frac{|PB|^2(K-c(\lambda_1))+J(\lambda_1)}{s(\lambda_1,\omega)^2}d\omega
\end{aligned}$$

Noting that $\Phi(\lambda_1) = J(\lambda_1)/(K-c(\lambda_1))$, the above can be re-written as

$$\begin{aligned}
\frac{d\Phi(\lambda_1)}{d\lambda_1} = \beta(\lambda_1)(K-c(\lambda_1))\Bigg[ &\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\frac{1}{|PB|^2+\lambda_1}d\omega\int_{-\pi}^{\pi}\frac{|PB|^2+\Phi(\lambda_1)}{|PB|^2+\lambda_1}d\omega \\
&- \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|PB|^2+\Phi(\lambda_1)}{(|PB|^2+\lambda_1)^2}d\omega\Bigg] \qquad (3.106)
\end{aligned}$$

Application of Theorem 3.18 to (3.106) leads directly to the conclusion that

$$\frac{d\Phi(\lambda_1)}{d\lambda_1} = 0 \iff \Phi(\lambda_1) = \lambda_1 \qquad (3.107a)$$

$$\frac{d\Phi(\lambda_1)}{d\lambda_1} < 0 \iff \Phi(\lambda_1) > \lambda_1 \qquad (3.107b)$$

$$\frac{d\Phi(\lambda_1)}{d\lambda_1} > 0 \iff \Phi(\lambda_1) < \lambda_1 \qquad (3.107c)$$

Substituting (3.105) into the right hand side of (3.107a) yields

$$\begin{aligned}
\lambda_1 K &= \beta_1(\lambda_1)\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|PB|^2}{|PB|^2+\lambda_1}d\omega + \lambda_1 c(\lambda_1) \\
&= \beta_1(\lambda_1)\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|PB|^2}{|PB|^2+\lambda_1}d\omega + \lambda_1\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{|PB|^2+\lambda_1}d\omega\right] = \beta(\lambda_1)
\end{aligned}$$

Thus,

$$\frac{d\Phi(\lambda_1)}{d\lambda_1} = 0 \iff K = k(\lambda_1) \triangleq \frac{\beta(\lambda_1)}{\lambda_1} = \frac{\mathrm{e}^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln(|PB|^2+\lambda_1)d\omega}}{\lambda_1} \tag{3.108}$$

The function $k(\lambda_1)$ is monotonically decreasing for all $\lambda_1 > 0$, since

$$\frac{dk(\lambda_1)}{d\lambda_1} = \frac{\beta(\lambda_1)}{\lambda_1^2}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\lambda_1}{|PB|^2+\lambda_1}d\omega - 1\right) = -\frac{\beta(\lambda_1)}{\lambda_1^2}\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|PB|^2}{|PB|^2+\lambda_1}d\omega < 0.$$

Therefore, the value of $\lambda_1$ that satisfies the right hand side of (3.108) (and yields $d\Phi(\lambda_1)/d\lambda_1 = 0$) is unique. From this and the fact that $\Phi(\lambda_1)$ and its derivative are continuous functions, together with (3.107), we conclude that the value of $\lambda_1$ that satisfies the right hand side of (3.108) is the unique minimizer of $\Phi(\lambda_1)$. This implies that

$$\min_f \mathscr{V}(f) = \min_{\lambda_1} \Phi(\lambda_1) = \lambda_1, \tag{3.109}$$

which substituted into (3.103) and (3.72) yields (3.89c). It also implies that the optimal $f$ is given by (3.95). In these solutions, $\lambda_1$ takes the unique value that satisfies the right hand side of (3.108). The latter equation is precisely (3.89b). In turn, (3.89a) is obtained by substituting (3.89b) into (3.95). This completes the proof. □

## 3.7 $B(z)$ Given

If only the post-filter $B(z)$ is given, then the minimization of the WCMSE reduces to the following

**Optimization Problem 3.6.** *For any given $K > 1$ and transfer function $B(z)$, find the frequency response $A(\mathrm{e}^{j\omega})$ and the frequency response magnitude $f(\mathrm{e}^{j\omega})$ that*

$$\textit{minimize}: \; D_{a,b}(\mathrm{x},\mathrm{y}) = a\frac{\|\Omega_{\mathrm{x}}A\|^2\|PBf\|^2}{K-\|f\|^2} + b\|(AB-1)\Omega_{\mathrm{x}}P\|^2 \tag{3.110a}$$

$$\textit{subject to}: \; 0 \le \frac{1}{2\pi}\int_{-\pi}^{\pi}\ln f(\mathrm{e}^{j\omega})d\omega, \tag{3.110b}$$

$$\|f\|^2 < K, \tag{3.110c}$$

$$f(\mathrm{e}^{j\omega}) \ge 0, \quad \forall\omega \in [-\pi,\pi]. \tag{3.110d}$$

**Theorem 3.8.** *The solution to Optimization Problem 3.6 is*

$$A(\mathrm{e}^{j\omega}) = \frac{b\left|P(\mathrm{e}^{j\omega})\right|^2 B(\mathrm{e}^{j\omega})^*}{a\lambda + b\left|P(\mathrm{e}^{j\omega})\right|^2\left|B(\mathrm{e}^{j\omega})\right|^2} \tag{3.111a}$$

$$f(\mathrm{e}^{j\omega})^2 = \frac{K\lambda}{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2+\lambda}, \tag{3.111b}$$

*where $\lambda$ is the unique scalar that satisfies*

$$\ln(K) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{\lambda} + 1\right) d\omega. \tag{3.111c}$$

*The minimum $D_{a,b}$, achieved with $A(z)$ and $f(\mathrm{e}^{j\omega})$ as in* (3.111)*, is*

$$\min_{A(z),f(\mathrm{e}^{j\omega})} D_{a,b}(\mathrm{x},\mathrm{y}) = ab\lambda\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2}{a\lambda + b\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2} d\omega \tag{3.112}$$

▲

*Proof.* If $B(z)$ is given, then the optimal $f(\mathrm{e}^{j\omega})$ does not depend on $A(z)$ (see 3.110a). Furthermore, the optimal $f(\mathrm{e}^{j\omega})$ is given by (3.89a). On the other hand, the optimal $A(z)$ given $B(z)$ equals the optimal $A(z)$ given the same $B(z)$ and given $f(\mathrm{e}^{j\omega})$ is optimal for that $B(z)$. Thus, the optimal $A(z)$ satisfies (3.34) with $f(\mathrm{e}^{j\omega})$ as in (3.89a). Furthermore, from Theorem (3.7), the optimal $f(\mathrm{e}^{j\omega})$ given $B(z)$ is such that $V$ in (3.34b) equals $\lambda$ (see (3.109) and (3.103)). When substituted into (3.34a), this yields (3.111a). Also, replacing $V$ by $\lambda$ in (3.35) yields (3.112). This completes the proof. □

## 3.8 $H(z)$ (pre-filter) Given

Here we find the filters that minimize the frequency weighted WCMSE for a given quantizer SNR under the architectural limitation that the pre-filter is given and fixed. In this case, it is necessary to make a distinction between the two schemes shown in Fig. 3.2, as discussed already in Section 1.2.2. For the configuration corresponding to Fig. 3.2-(a), it is implicit that one can both measure and act upon the signal coming out of the pre-filter $A(z)$. As a consequence, even if $A(z)$ is fixed, one could alter the transfer function from $\{\mathrm{x}(k)\}$ to $\{\mathrm{v}(k)\}$ at will. Hence, assuming in this architecture that the pre-filter $A(z)$ is fixed and given makes little practical sense.

By contrast, the scenario in which the pre-filter is fixed and given is better represented by the configuration shown in Fig. 3.2-(b). This scheme assumes implicitly that one can add signals to the input of the scalar quantizer, but not measure the result of this addition. Hence, it is not possible to bypass the limitations imposed by a fixed $H(z)$, i.e., one cannot alter the transfer function from $\{\mathrm{x}(k)\}$ to $\{\mathrm{v}(k)\}$ without changing the transfer function from $\{\mathrm{n}(k)\}$ to $\{\mathrm{v}(k)\}$ also.

Focusing on the scheme of Fig. 3.2-(b), if only the pre-filter $H(z)$ is fixed and given, then the minimization of the WCMSE reduces to the following:

**Optimization Problem 3.7.** *For any given $K > 1$ and $H(z)$, find the filter $L(z)$ and the frequency response magnitude $f(\mathrm{e}^{j\omega})$ that*

$$\textit{minimize}: \; D_{a,b}(\mathrm{x},\mathrm{y}) = a\frac{\|PL\|^2\|\Omega_{\mathrm{x}}Hf\|^2}{K-\|f\|^2} + b\|(HL-1)\Omega_{\mathrm{x}}\|^2 \tag{3.113a}$$

$$\textit{subject to}: \; 0 \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(\mathrm{e}^{j\omega})d\omega, \tag{3.113b}$$

$$\|f\|^2 < K, \tag{3.113c}$$

$$f(\mathrm{e}^{j\omega}) \geq 0, \quad \forall\omega \in [-\pi,\pi]. \tag{3.113d}$$

**Theorem 3.9.** *The solution to Optimization Problem 3.7 is*

$$L(\mathrm{e}^{j\omega}) = \frac{b\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^2 H(\mathrm{e}^{j\omega})^*}{a\lambda + b\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^2\left|H(\mathrm{e}^{j\omega})\right|^2} \tag{3.114a}$$

$$f(\mathrm{e}^{j\omega})^2 = \frac{K\lambda}{\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})H(\mathrm{e}^{j\omega})\right|^2 + \lambda}, \tag{3.114b}$$

*where $\lambda$ is the unique scalar that satisfies*

$$\ln(K) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\frac{\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})H(\mathrm{e}^{j\omega})\right|^2}{\lambda} + 1\right)d\omega. \tag{3.114c}$$

*The minimum $D_{a,b}$, achieved with $L(z)$ and $f(\mathrm{e}^{j\omega})$ as in* (3.114)*, is*

$$\min_{L(z),f(\mathrm{e}^{j\omega})} D_{a,b}(\mathrm{x},\mathrm{y}) = ab\lambda\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2}{a\lambda + b\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})H(\mathrm{e}^{j\omega})\right|^2}d\omega \tag{3.115}$$

▲

*Proof.* The cost functional in (3.113a) has the same structure as that in Theorem 3.8. More precisely, the constraint of a fixed $H(z)$ in (3.113a) plays the same role as the constraint of a fixed $B(z)$ in (3.110a). Thus, the solution is given by (3.111), replacing $\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})$ by $P(\mathrm{e}^{j\omega})$, $B(\mathrm{e}^{j\omega})$ by $H(\mathrm{e}^{j\omega})$, and $A(\mathrm{e}^{j\omega})$ by $L(\mathrm{e}^{j\omega})$, which yields (3.114). This completes the proof. □

## 3.9 No Constraints: The WCMSE Optimal Feedback Quantizer

We now address the problem of finding the frequency responses of $A(z)$, $B(z)$ and $F(z)$ that minimize the WCMSE for a given quantizer SNR $\gamma = K-1$. These frequency responses characterize the WCMSE-optimal FQ when all three degrees of freedom are available.

**Optimization Problem 3.8.** *For any given $K > 1$, find the filters $A(z)$, $B(z)$ and the frequency response magnitude $f(\mathrm{e}^{j\omega})$ that*

$$\textit{minimize}: \ a\frac{\|\Omega_{\mathrm{x}}A\|^2\|PBf\|^2}{K-\|f\|^2} + b\|(AB-1)\Omega_{\mathrm{x}}\|^2 \tag{3.116a}$$

$$\textit{subject to}: \ 0 \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(\mathrm{e}^{j\omega})d\omega, \tag{3.116b}$$

$$\|f\|^2 < K, \tag{3.116c}$$

$$f(\mathrm{e}^{j\omega}) \geq 0, \quad \forall\omega \in [-\pi,\pi]. \tag{3.116d}$$

▲

The solution to Optimization Problem 3.8 is given in the next theorem, for the cases in which $a \leq 2b$. The latter restriction is imposed in order to avoid the high mathematical complexity that arises whenever $a > 2b$ and the distortion is larger than a given threshold, as explained in footnote 9 of the proof. We shall say more about the implications of the relationship $a > 2b$ later, when we argue, in in Sections 4.3.2 and 4.7, that the condition $a \leq 2b$ is satisfied in many cases of practical interest.

**Theorem 3.10.** *If $b/a \geq 2$, the solution to Optimization Problem 3.8 is*

$$f(\mathrm{e}^{j\omega}) = \frac{\sqrt{K\alpha}}{\sqrt{G(\mathrm{e}^{j\omega})^2 + \left[1-\frac{a}{b}\right]\alpha} + G(\mathrm{e}^{j\omega})}, \tag{3.117a}$$

$$A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega}) = 1 - \frac{(a/b)\alpha/2}{\left(\sqrt{G(\mathrm{e}^{j\omega})^2 + \left[1-\frac{a}{b}\right]\alpha} + G(\mathrm{e}^{j\omega})\right)G(\mathrm{e}^{j\omega})}, \tag{3.117b}$$

$$\left|B(\mathrm{e}^{j\omega})\right| = \frac{\left|P(\mathrm{e}^{j\omega})\right|^{\sim 1}}{\kappa}\sqrt{\frac{\left(\sqrt{G(\mathrm{e}^{j\omega})^2 + \left[1-\frac{a}{b}\right]\alpha} + G(\mathrm{e}^{j\omega})\right)^2 - \alpha}{2\sqrt{K\alpha}}}, \tag{3.117c}$$

$$\left|A(\mathrm{e}^{j\omega})\right| = \kappa\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^{\sim 1}\sqrt{\frac{\sqrt{K\alpha}}{2}\left[1 - \frac{\alpha}{\left(\sqrt{G(\mathrm{e}^{j\omega})^2 + \left[1-\frac{a}{b}\right]\alpha} + G(\mathrm{e}^{j\omega})\right)^2}\right]}, \tag{3.117d}$$

*a.e. on $[-\pi,\pi]$, where $\alpha > 0$ is the unique scalar that satisfies*

$$\frac{1}{2}\ln(K) = \frac{1}{2\pi}\int\limits_{|\Omega_{\mathrm{x}}P| \geq \frac{a\sqrt{\alpha}}{2b}} \ln\left(\sqrt{\frac{\left|P(\mathrm{e}^{j\omega})\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^2}{\alpha} + \left[1-\frac{a}{b}\right]} + \frac{\left|P(\mathrm{e}^{j\omega})\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|}{\sqrt{\alpha}}\right)d\omega \tag{3.118}$$

*and*

$$G(\mathrm{e}^{j\omega}) \triangleq \max\left\{\frac{a}{b}\sqrt{\alpha}/2\ ,\ \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|\right\}, \quad \forall\omega \in [-\pi,\pi]. \tag{3.119}$$

*The scalar $\alpha$ is related to $K$, $\kappa$ and the variance $\sigma_{\mathrm{n}}^2$ via*

$$\sigma_{\mathrm{n}}^2 = \frac{\kappa^2}{2}\sqrt{\alpha/K} \tag{3.120}$$

*The minimum $D_{a,b}(\mathrm{x},\mathrm{y})$, achieved with* (3.117)*, is*

$$\min_{A(z),B(z),f(\mathrm{e}^{j\omega})} D_{a,b}(\mathrm{x},\mathrm{y}) = \frac{a}{2\pi}\int\limits_{|\Omega_{\mathrm{x}}P|\geq\frac{a\sqrt{\alpha}}{2b}} \frac{(\alpha/2)\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|}{\sqrt{\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2 + [1-\frac{a}{b}]\alpha} + \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|} d\omega + \frac{b}{2\pi}\int\limits_{|\Omega_{\mathrm{x}}P|<\frac{a\sqrt{\alpha}}{2b}} \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2 d\omega \tag{3.121}$$

▲

*Proof.* Clearly, the optimal filters must also be reciprocally optimal. In particular, Theorem 3.7 must hold. This implies that $\|PBf\|^2/(K-\|f\|^2) = \lambda$, see (3.109) and (3.103). Substituting the latter into (3.54), we have that

$$\xi = \frac{a}{b}\cdot\frac{\|\Omega_{\mathrm{x}}A\|^2}{\kappa^2\|PBf\|^2}\lambda. \tag{3.122}$$

Since the optimal filters also satisfy Theorem 3.3 (page 56), we have from (3.41) that $\kappa^2\|PBf\|^2 = \frac{1}{\kappa^2}\|\Omega_{\mathrm{x}}A\|^2$. Substitution of the latter into (3.122) yields

$$\xi = \frac{a}{b}\kappa^2\lambda = (1/2)\tfrac{a}{b}\sqrt{\alpha/K}. \tag{3.123}$$

where

$$\alpha \triangleq 4K\kappa^4\lambda^2 = 4K\left(\tfrac{b}{a}\right)^2\xi^2. \tag{3.124}$$

If the filters are reciprocally optimal, then (3.51a), (3.89a) and (3.89b) hold simultaneously. In particular, from (3.89a),

$$f(\mathrm{e}^{j\omega})^2 = \frac{K\lambda}{\left|P(\mathrm{e}^{j\omega})\right|^2\left|B(\mathrm{e}^{j\omega})\right|^2+\lambda} \tag{3.125}$$

From (3.53) and (3.51a) it follows that $\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right| = 0 \iff G(\mathrm{e}^{j\omega}) = \xi f(\mathrm{e}^{j\omega}) \iff \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| \leq \xi f(\mathrm{e}^{j\omega})$, for any given frequency $\omega\in[-\pi,\pi]$. On the other hand, from (3.125) we have that $\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right| = 0 \iff f(\mathrm{e}^{j\omega}) = \sqrt{K}$. Thus

$$\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| \leq \xi f(\mathrm{e}^{j\omega}) \iff \xi f(\mathrm{e}^{j\omega}) = \xi\sqrt{K} = \frac{a}{b}\cdot\frac{\sqrt{\alpha}}{2}. \tag{3.126}$$

Substitution of (3.126) into (3.53) yields (3.119). Substitution of (3.123) into (3.54) leads directly to (3.120).

On the other hand, substitution of (3.51a) into (3.125) yields

$$f(\mathrm{e}^{j\omega})^2 = \frac{K\lambda}{\frac{1}{\kappa^2}\left(\frac{G(\mathrm{e}^{j\omega})}{f(\mathrm{e}^{j\omega})} - \xi\right) + \lambda} = \frac{K\kappa^2\lambda}{\frac{G(\mathrm{e}^{j\omega})}{f(\mathrm{e}^{j\omega})} + \kappa^2\lambda - \xi} \iff$$
$$0 = [\kappa^2\lambda - \xi]f(\mathrm{e}^{j\omega})^2 + G(\mathrm{e}^{j\omega})f(\mathrm{e}^{j\omega}) - K\kappa^2\lambda,$$

where we recall from (3.53) that

$$G(\mathrm{e}^{j\omega}) = \max\left\{\xi f(\mathrm{e}^{j\omega}), |\Omega_{\mathrm{x}}P|\right\}, \quad \forall\omega \in [-\pi, \pi]. \tag{3.127}$$

If $\kappa^2\lambda - \xi = 0$, then the optimal $f(\mathrm{e}^{j\omega})$ is

$$f(\mathrm{e}^{j\omega}) = \frac{K\kappa^2\lambda}{G(\mathrm{e}^{j\omega})} = \frac{\sqrt{K\alpha/4}}{G(\mathrm{e}^{j\omega})}, \quad \text{a.e. on } [-\pi, \pi]. \tag{3.128}$$

Otherwise, if $\kappa^2\lambda - \xi \neq 0$,

$$f(\mathrm{e}^{j\omega}) = \frac{\pm\sqrt{G(\mathrm{e}^{j\omega})^2 + 4K\kappa^2\lambda[\kappa^2\lambda - \xi]} - G(\mathrm{e}^{j\omega})}{2[\kappa^2\lambda - \xi]} = \frac{\pm\sqrt{G(\mathrm{e}^{j\omega})^2 + 4K\kappa^2\lambda[\kappa^2\lambda - \frac{a}{b}\kappa^2\lambda]} - G(\mathrm{e}^{j\omega})}{2[\kappa^2\lambda - \frac{a}{b}\kappa^2\lambda]}$$
$$= \frac{\pm\sqrt{G(\mathrm{e}^{j\omega})^2 + 4K\kappa^4\lambda^2[1 - \frac{a}{b}]} - G(\mathrm{e}^{j\omega})}{2\kappa^2\lambda[1 - \frac{a}{b}]} = \frac{\pm\sqrt{G(\mathrm{e}^{j\omega})^2 + [1 - \frac{a}{b}]\alpha} - G(\mathrm{e}^{j\omega})}{[1 - \frac{a}{b}]\sqrt{\alpha/K}} \tag{3.129}$$

It is now necessary to determine which sign before the square root in (3.129) yields the solution. We will next show that the minus sign ($-\sqrt{}$) yields an infeasible solution. To this end, we note from (3.127) that a feasible solution $f(\mathrm{e}^{j\omega})$ must satisfy the condition

$$0 \leq \frac{\xi f(\mathrm{e}^{j\omega})}{G(\mathrm{e}^{j\omega})} \leq 1 \tag{3.130}$$

for all $\omega \in [-\pi, \pi]$. Substituting (3.123) and (3.129) with the choice $-\sqrt{}$ into (3.130), the latter condition becomes

$$\frac{\xi f(\mathrm{e}^{j\omega})}{G(\mathrm{e}^{j\omega})} = \frac{1 + \sqrt{1 - \frac{[\frac{a}{b} - 1]\alpha}{G(\mathrm{e}^{j\omega})^2}}}{2\frac{b}{a}[\frac{a}{b} - 1]}, \tag{3.131}$$

from where it follows immediately that the choice $-\sqrt{}$ is infeasible if $a < b$. On the other hand, when $a > b$, the right-hand side of (3.131) increases monotonically with $G$. Since, from (3.126) and (3.127) $G(\mathrm{e}^{j\omega}) \geq \frac{a}{2b}\sqrt{\alpha},\ \forall\omega \in [-\pi, \pi]$, we have from (3.131) that

$$\frac{\xi f(\mathrm{e}^{j\omega})}{G(\mathrm{e}^{j\omega})} \geq \frac{1 + \sqrt{1 - \frac{[\frac{a}{b} - 1]\alpha}{(\frac{a}{2b}\sqrt{\alpha})^2}}}{2\frac{b}{a}[\frac{a}{b} - 1]} = \frac{1 + \frac{2b}{a}\left|\frac{a}{2b} - 1\right|}{2 - \frac{2b}{a}} = \frac{1 + \left|1 - \frac{2b}{a}\right|}{1 + 1 - \frac{2b}{a}} \geq 1. \tag{3.132}$$

The last inequality is strict if $b < a < 2b$, and becomes equality only if $a \geq 2b$. Thus, when $a < 2b$, choosing the minus sign before the squared root in (3.129) leads to an infeasible solution for all $\omega \in$

$[-\pi, \pi]$. On the other hand, if $a = 2b$, then the minus sign before the squared root in (3.129) becomes feasible only for frequencies at which $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| \leq \frac{a}{2b}\sqrt{\alpha}$. However, the solution obtained with the plus before the squared root in (3.129) yields the same result over those frequencies.[9] With the choice $+\sqrt{\ }$ in (3.129), the latter becomes (3.117a). Notice that this solution also yields (3.128) if $\kappa^2\lambda - \xi = 0$ (which happens if and only if $a = b$). In addition, noting from (3.125) that

$$(\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 + \lambda)/\lambda = K/f(\mathrm{e}^{j\omega})^2 = \frac{\left(\sqrt{G(\mathrm{e}^{j\omega})^2 + [1 - \frac{a}{b}]\alpha} + G(\mathrm{e}^{j\omega})\right)^2}{\alpha},$$

and substituting this into (3.89b), we conclude that $\alpha$ is the unique scalar that satisfies

$$\frac{1}{2}\ln(K) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\frac{\sqrt{G(\mathrm{e}^{j\omega})^2 + [1 - \frac{a}{b}]\alpha} + G(\mathrm{e}^{j\omega})}{\sqrt{\alpha}}\right) d\omega. \tag{3.133}$$

We have that (3.125) is precisely (3.117a), and that (3.119) together with (3.133) lead directly to (3.118). Substitution of (3.117a) into (3.55) yields (3.121). Finally, substitution of (3.125) and (3.123) into (3.51) yields (3.117c) and (3.117d). This completes the proof. □

### 3.9.1 Special Cases

By using Theorem 3.10, it is possible to characterize the optimal filters and SNR-distortion performance of optimal feedback quantizers for each possible combination of values for the weights $a$, $b$. Two relevant special cases are discussed below.

#### MSE-Optimal Feedback Quantization

If one sets $a = b = 1$, then WCMSE is equivalent to standard MSE. In this case, Theorem 3.10 yields that, for an MSE-optimal feedback quantizer,

$$f(\mathrm{e}^{j\omega}) = \frac{\sqrt{K\alpha/4}}{G(\mathrm{e}^{j\omega})}, \tag{3.134a}$$

$$A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega}) = 1 - \frac{\alpha/4}{G(\mathrm{e}^{j\omega})^2}, \tag{3.134b}$$

$$\left|B(\mathrm{e}^{j\omega})\right| = \frac{\left|P(\mathrm{e}^{j\omega})\right|^{\sim 1}}{\kappa}\sqrt{\frac{G(\mathrm{e}^{j\omega})^2 - \alpha/4}{\sqrt{K\alpha/4}}}, \tag{3.134c}$$

$$\left|A(\mathrm{e}^{j\omega})\right| = \kappa\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^{\sim 1}\sqrt{\sqrt{K\alpha/4}\left[1 - \frac{\alpha/4}{G(\mathrm{e}^{j\omega})^2}\right]} \tag{3.134d}$$

[9] On the other hand, if $a > 2b$, both choices of sign lead to a feasible over all frequencies $\omega$ at which $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| < \frac{a}{2b}\sqrt{\alpha}$. This difficulty is avoided by excluding the cases $a > 2b$ from the analysis.

a.e. on $[-\pi, \pi]$, where $\alpha > 0$ is the unique scalar that satisfies

$$\frac{1}{2}\ln(K) = \frac{1}{2\pi} \int\limits_{|\Omega_{\mathrm{x}} P| \geq \sqrt{\alpha/4}} \ln\left( \frac{\left|P(\mathrm{e}^{j\omega})\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|}{\sqrt{\alpha/4}} \right) d\omega \tag{3.135}$$

and

$$G(\mathrm{e}^{j\omega}) \triangleq \max\left\{ \sqrt{\alpha/4}\ ,\ \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| \right\}, \quad \forall \omega \in [-\pi, \pi]. \tag{3.136}$$

The minimum $D_{1,1}(\mathrm{x}, \mathrm{y})$, achieved with (3.134), is

$$\min_{A(z),B(z),f(\mathrm{e}^{j\omega})} D_{1,1}(\mathrm{x}, \mathrm{y}) = \frac{1}{2\pi} \int\limits_{-\pi}^{\pi} \min\left\{ \frac{\alpha}{4}\ ,\ \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2 \right\} d\omega \tag{3.137}$$

Notice that if $\left|P(\mathrm{e}^{j\omega})\right| \equiv 1$, and with the change of variable

$$\theta = \alpha/4, \tag{3.138}$$

the expression for the MSE in (3.137) is equivalent to the one given by the water-filling equations (1.1). Moreover, the filters characterized by (3.134) are equivalent to the filters that achieve the quadratic Gaussian rate distortion function in [15]. Notice also that, with the change of variable (3.138), the quantity $\frac{1}{2}\ln(K)$ in (3.135) plays the role of the rate $R(D)$ in (1.1). In Chapter 5 we shall see that this correspondence is not accidental, and that it has important implications in the design of optimal ED pairs.

**Optimal Perfect Reconstruction Feedback Quantization**

If one sets $a = 1$ and lets $b \to \infty$, then from (3.117b) the optimal filters $A(z)$ and $B(z)$ satisfy the perfect reconstruction condition

$$A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega}) = 1, \ \ \forall \omega \text{ such that } S_x(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega}) \neq 0. \tag{3.139}$$

As a consequence, the WCMSE in this case is made of source-uncorrelated reconstruction error only. From (3.117), the optimal frequency responses for $a = 1$ and $b = \infty$ are found to be:

$$f(\mathrm{e}^{j\omega}) = \frac{\sqrt{K\alpha}}{\sqrt{g(\omega)^2 + \alpha} + g(\omega)}, \tag{3.140a}$$

$$\left|B(\mathrm{e}^{j\omega})\right| = \frac{\left|P(\mathrm{e}^{j\omega})\right|^{\sim 1}}{\kappa} \sqrt{\frac{g(\omega)\left(\sqrt{g(\omega)^2 + \alpha} + g(\omega)\right)}{\sqrt{K\alpha}}}, \tag{3.140b}$$

$$\left|A(\mathrm{e}^{j\omega})\right| = \kappa \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^{\sim 1} \sqrt{\frac{\sqrt{K\alpha}\, g(\omega)}{\sqrt{g(\omega)^2 + \alpha} + g(\omega)}} \tag{3.140c}$$

a.e. on $[-\pi, \pi]$, where

$$g(\omega) \triangleq \left|P(\mathrm{e}^{j\omega})\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|, \quad \forall \omega \in [-\pi, \pi], \tag{3.141}$$

and where $\alpha > 0$ is the unique scalar that satisfies

$$\frac{1}{2}\ln(K) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\sqrt{\frac{g(\omega)^2}{\alpha}+1} \; + \; \frac{g(\omega)}{\sqrt{\alpha}}\right) d\omega. \tag{3.142}$$

The minimum $D_{1,\infty}(\mathrm{x},\mathrm{y})$, achieved with (3.117), is

$$\min_{A(z),B(z),f(\mathrm{e}^{j\omega})} D_{1,\infty}(\mathrm{x},\mathrm{y}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{(\alpha/2)g(\omega)}{\sqrt{g(\omega)^2+\alpha}+g(\omega)} d\omega \tag{3.143}$$

Recalling that $\kappa$ is an arbitrary scalar, (3.117) is exactly the solution for optimal perfect reconstruction feedback quantizers first derived by the author and colleagues in [118].

It can be shown that the right-hand side of (3.142) is a convex, monotonically decreasing function of $\alpha$. This guarantees that, for any given $K$, the value of $\alpha$ satisfying (3.142) can be easily found via, for example, the bisection algorithm [146], or, indeed, by any other convex optimization method [147].

It is interesting to note that for the perfect reconstruction case, we have

$$K \to 1 \iff \alpha \to \infty. \tag{3.144}$$

In such a case, it can be seen from (3.140a) that the optimal noise-shaping frequency response magnitude $f(\mathrm{e}^{j\omega})$ converges uniformly to 1, i.e., the feedback filer $F(z)$ approaches 0. This situation corresponds to not using feedback. In view of (3.140), expression (3.144) also implies that when $K \to 1$, $A(z)$ and $B(z)$ converge to *half-whitening filters*, which are known to be the best perfect reconstruction pre/post-filters in the absence of feedback, see, e.g., [55, 81].

An important feature of the filters characterized in (3.117) is that they all can be implemented with arbitrary accuracy by using *causal* filters (see also lemmas 3.6 and 3.4, in pages 65 and 57, respectively). This is not only attractive in applications wherein there exists feedback between reconstruction and source, but will also be instrumental in our derivation of the bounds for the causal rate-distortion function in Chapter 6.

### 3.9.2 The Importance of Taking Account of Fed Back Quantization Noise

If one tried to optimize the filters of a FQ neglecting fed back quantization noise, i.e., by trying to minimize $a\frac{\|A\Omega_x\|^2\|(1-F)BP\|^2}{\gamma} + b\|(AB-1)\Omega_{\mathrm{x}}P\|^2$ (compare to (3.23)), then one would obtain a (sub optimal) feedback filter, namely $F_0(z)$, which satisfies

$$|1 - F_0| = \eta_{\Omega_{\mathrm{x}}P}\,|\Omega_{\mathrm{x}}P|^{-1}, \tag{3.145a}$$

where the minimal prediction variance $\eta_{\Omega_{\rm x}P}$, already defined in (3.70) (page 63), is given by

$$\eta_{\Omega_{\rm x}P} = {\rm e}^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|\Omega_{\rm x}({\rm e}^{j\omega})P({\rm e}^{j\omega})\right| d\omega}, \tag{3.145b}$$

provided $|\Omega_{\rm x}P| > 0,\ \forall\omega \in [-\pi,\pi]$. This is the solution one obtains by setting $K \to \infty$ in (3.117a) and (3.118). It corresponds to the result obtained in [81], which was restricted to the cases where $\gamma \gg \|F_0\|^2$. For the case $\Omega_{\rm x}({\rm e}^{j\omega}) \equiv 1$, the noise transfer function magnitude $|1 - F_0(z)|$ is also equivalent to that derived in [88]. The latter is optimal in the sense of minimizing the ratio $\sigma_\epsilon^2/\sigma_n^2$, but not in the sense of minimizing $\sigma_\epsilon^2$ for a fixed quantizer SNR $\gamma$.

As it can be seen from (3.117a) and (3.118), $f^\star$, in general, does indeed approach $f_0 = |1 - F_0|$ as $\gamma \to \infty$. Hence, one can expect $F_0$ to be *near* optimal in situations where $\gamma \gg \|F_0\|^2$, see (3.23). The latter is often satisfied at high bit-rates (i.e., when many quantization levels are available). However, for any given number of quantization levels, it is easy to find practical situations where $\Omega_{\rm x}P$ is such that $\|F_0\|^2$ is comparable to (or greater than) $\gamma$. To see this, suppose that there exist scalars $c > 1$, $\ell > 0$ and $H \geq c\,\ell$ such that

$$\left|\Omega_{\rm x}({\rm e}^{j\omega})P({\rm e}^{j\omega})\right| \geq \ell, \quad \forall\omega \in [-\pi,\pi], \tag{3.146}$$

$$\left|\Omega_{\rm x}({\rm e}^{j\omega})P({\rm e}^{j\omega})\right| \leq c\,\ell, \quad \forall\omega \in \mathbb{L}, \tag{3.147}$$

$$\left|\Omega_{\rm x}({\rm e}^{j\omega})P({\rm e}^{j\omega})\right| \geq H, \quad \forall\omega \in \mathbb{H}, \tag{3.148}$$

where $\mathbb{L}$ and $\mathbb{H}$ are subsets of $[-\pi,\pi]$ having Lebesgue measures $|\mathbb{L}|$ and $|\mathbb{H}|$, respectively. Then

$$\begin{aligned}
\|1 - F_0\|^2 &= \exp\left(\frac{1}{2\pi}\int\limits_{\omega\in\mathbb{H}} \ln\left|\Omega_{\rm x}({\rm e}^{j\omega})P({\rm e}^{j\omega})\right|^2 d\omega + \frac{1}{2\pi}\int\limits_{\omega\notin\mathbb{H}} \ln\left|\Omega_{\rm x}({\rm e}^{j\omega})P({\rm e}^{j\omega})\right|^2 d\omega\right) \\
&\qquad \times \left[\frac{1}{2\pi}\int\limits_{\omega\notin\mathbb{L}} \frac{d\omega}{\left|\Omega_{\rm x}({\rm e}^{j\omega})P({\rm e}^{j\omega})\right|^2} + \frac{1}{2\pi}\int\limits_{\omega\in\mathbb{L}} \frac{d\omega}{\left|\Omega_{\rm x}({\rm e}^{j\omega})P({\rm e}^{j\omega})\right|^2}\right] \\
&\geq \exp\left(\frac{|\mathbb{H}|}{2\pi}\ln H^2 + \frac{2\pi - |\mathbb{H}|}{2\pi}\ln\ell^2\right)\left[\frac{|\mathbb{L}|}{2\pi}(c^2\ell^2)^{-1}\right] \\
&= \frac{|\mathbb{L}|}{2\pi c^2}\left(H^2\right)^{\frac{|\mathbb{H}|}{2\pi}}\left(\ell^2\right)^{-\frac{|\mathbb{H}|}{2\pi}} = \frac{|\mathbb{L}|}{2\pi c^2}\left(\frac{H}{\ell}\right)^{\frac{|\mathbb{H}|}{\pi}}
\end{aligned}$$

Recalling that $\|F - 1\|^2 = \|F\|^2 + 1$ (see (3.19)), the above yields

$$\|F_0\|^2 \geq \frac{|\mathbb{L}|}{2\pi c^2}\left(\frac{H}{\ell}\right)^{|\mathbb{H}|/\pi}. \tag{3.149}$$

This implies that *a large $\|F_0\|^2$ is obtained for any product $\Omega_{\rm x}P$ whose magnitude becomes significantly small (in relative terms) over certain frequency bands.* (An example is included in Section 3.11 below.) A direct consequence is that, for these cases, and in view of (3.23), trying to match $|1 - F|$ to

$\eta_{\Omega_{\mathrm{x}}P}\,|\Omega_{\mathrm{x}}P|^{-1}$ will yield a performance far from optimal. Also, this also increases the risk of incurring large limit-cycle oscillations if no clipping is employed (see, e.g., [43, 78]).

The (possibly unbounded) increase of $\|F\|^2$ as $|1-F|$ approaches $\eta_{\Omega_{\mathrm{x}}P}\,|\Omega_{\mathrm{x}}P|^{-1}$ seems to have been first observed in [79]. Several heuristic solutions have been proposed since then (see, e.g., [43, 46, 78, 82, 83, 88]). In contrast to these approaches, the method derived in the present work allows one to characterize the true optimal filters, by explicitly taking into account $\|F\|^2$ in the cost functional to be minimized (see (3.23)). In other words, our method not only guarantees that $\|F\|^2 < \gamma$, but also yields the true optimal filters. Our proposal also has the advantage of being applicable to arbitrary input spectra and frequency weighting functions, regardless of how small the quantizer SNR $\gamma$ may be, within the scope of validity of the Linear Model.

## 3.10 Comparative Analysis

### 3.10.1 Optimal Frequency Responses

The frequency responses of the optimal filters $A(z)$, $B(z)$ and $1 - F(z)$ for each architecturally constrained scenario are listed in Table 3.1. It is interesting to note that, in all the cases, the optimal frequency response magnitudes $\left|A(\mathrm{e}^{j\omega})\right|$, $\left|B(\mathrm{e}^{j\omega})\right|$, $f(\mathrm{e}^{j\omega})$ are, in general, different from unity, unless $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$ is constant over $[-\pi, \pi]$. This implies that, unless the frequency weighted input spectrum has flat PSD, it is always necessary to utilize all the available degrees of freedom in order to achieve optimal performance, in each scenario. This fact goes against the intuitive idea that, in the scheme shown in Fig 3.1, the filter $F(z)$ is necessary only when error frequency weighting makes noise-shaping a useful tool to reduce reconstruction error. It also contradicts the perhaps natural thinking that the pre-filter $A(z)$ is required only when the input has a non-flat spectrum that gives room for predictive pre-filtering. In reality, and as can be seen from Table 3.1, pre-filtering is also beneficial when $\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})$ is constant and $P(\mathrm{e}^{j\omega})$ is not, while noise-shaping is required for optimality, even when $P(\mathrm{e}^{j\omega})$ is constant, as long as $\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})$ is not. More generally, it is clear that, unless $\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$ is constant, every degree of freedom not available, or not exploited, in the design of a feedback quantizer, will always entail a penalty in operational rate-distortion performance.

### 3.10.2 Optimal Signal Spectra

Here we will analyze the PSD of the output of the quantizer, $S_{\mathrm{w}}(\mathrm{e}^{j\omega})$, and the PSD of the frequency weighted reconstruction error, $S_{\epsilon}(\mathrm{e}^{j\omega})$. Table 3.2 lists the expressions for the optimal $S_{\mathrm{w}}(\mathrm{e}^{j\omega})$ in each scenario, derived from the equations characterizing the optimal filter frequency responses. It can be seen from Table 3.2 that, unless $\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})$ and $P(\mathrm{e}^{j\omega})$ and the frequency responses assumed given in each scenario take special forms, $S_{\mathrm{w}}(\mathrm{e}^{j\omega})$ is, in general, not constant. However, when all three degrees of freedom are available (last row in Table 3.2), $S_{\mathrm{w}}(\mathrm{e}^{j\omega})$ is constant *for any input spectral density, frequency weighting criterion, and choice of* $a$, $b$. Having a flat PSD in the output of $\mathcal{Q}$ is beneficial, since it allows one to achieve near optimal coding of the quantizer output with a memory-less entropy coder (see (2.60) on page 42 and Lemma 5.2 in Section 5.2). Conversely, if any of the three degrees of freedom is not utilized optimally, then rate-distortion performance can be improved by using entropy coding with memory. Indeed, it will be shown in Chapter 5 (Section 5.2.2) that, when using subtractively dithered scalar quantization, entropy coding with infinite memory is capable of substituting the lack of any (but not more than one) of the three degrees of freedom associated with an FQ scheme.

Table 3.2 also summarizes the expressions for the PSD of the frequency weighted reconstruction

**Table 3.1:** Optimal frequency response magnitudes

| Given | Optimal $A(z)$ | Optimal $B(z)$ | Optimal $f$ | Where |
|---|---|---|---|---|
| $A$, $f$ | - | $\frac{b\lvert\Omega_{\mathrm{x}}\rvert^2 A^*}{a\sigma_v^2 f^2+b\lvert\Omega_{\mathrm{x}}A\rvert^2}$ | - | - |
| $B$, $f$ | $\frac{b\lvert P\rvert^2 B^*}{aV+b\lvert PB\rvert^2}$ | - | - | (i) |
| $f$, $AB=W$ | $\kappa\sqrt{\lvert P\rvert\lvert\Omega_{\mathrm{x}}\rvert^{\sim 1} f\lvert W\rvert}$ | $\frac{1}{\kappa}\sqrt{\lvert P\rvert^{\sim 1}\lvert\Omega_{\mathrm{x}}\rvert f^{\sim 1}\lvert W\rvert}$ | - | - |
| $f$ | $\kappa\lvert\Omega_{\mathrm{x}}\rvert^{\sim 1}\sqrt{Gf-\xi f^2}$ | $\frac{\lvert P\rvert^{\sim 1}}{\kappa}\sqrt{\frac{G}{f}-\xi}$ | - | (ii) |
| $A$, $B$ | - | - | $\sqrt{\frac{K\lambda}{\lvert PB\rvert^2+\lambda}}$ | (iii) |
| $B$ | $\frac{b\lvert P\rvert^2 B^*}{a\lambda+b\lvert PB\rvert^2}$ | - | $\sqrt{\frac{K\lambda}{\lvert PB\rvert^2+\lambda}}$ | (iv) |
| $H$ | - | $L=\frac{b\lvert\Omega_{\mathrm{x}}\rvert^2 H^*}{a\lambda+b\lvert\Omega_{\mathrm{x}}H\rvert^2}$ | $\sqrt{\frac{K\lambda}{\lvert\Omega_{\mathrm{x}}H\rvert^2+\lambda}}$ | (v) |
| - | $\kappa\lvert\Omega_{\mathrm{x}}\rvert^{\sim 1}\sqrt{\frac{\sqrt{K\alpha}}{2}\left[1-\frac{\alpha}{\left(\sqrt{G^2+\zeta\alpha}+G\right)^2}\right]}$ | $\frac{\lvert P\rvert^{\sim 1}}{\kappa}\sqrt{\frac{\left(\sqrt{G^2+\zeta\alpha}+G\right)^2-\alpha}{2\sqrt{K\alpha}}}$ | $\frac{\sqrt{K\alpha}}{\sqrt{G^2+\zeta\alpha}+G}$ | (vi) |

(i): $V \triangleq \frac{\|PBf\|^2}{K-\|f\|^2}$

(ii): $\kappa > 0$ is an arbitrary real constant, $\xi$ is the unique scalar that satisfies

$K = \frac{a}{b}\frac{1}{2\pi}\int_{-\pi}^{\pi} \max\left\{f(\mathrm{e}^{j\omega})^2\ ,\ \frac{f(\mathrm{e}^{j\omega})\left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|}{\xi}\right\}d\omega + [1-\frac{a}{b}]\|f\|^2$, and

$G(\mathrm{e}^{j\omega}) \triangleq \max\left\{\xi f(\mathrm{e}^{j\omega})\ ,\ \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|\right\}, \quad \forall\omega\in[-\pi,\pi]$

(iii): $\ln(K) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{\lambda}+1\right)d\omega.$

(iv): $\ln(K) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{\lambda}+1\right)d\omega.$

(v): $\ln(K) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\frac{\lambda_2}{\lambda_2-\lvert\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})A(\mathrm{e}^{j\omega})\rvert^2}\right)d\omega$

(vi): $\frac{1}{2}\ln(K) = \frac{1}{2\pi}\int\limits_{|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})|\geq\frac{a\sqrt{\alpha}}{2b}} \ln\left(\sqrt{\frac{|P(\mathrm{e}^{j\omega})\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})|^2}{\alpha}+\zeta}\ +\ \frac{\left|P(\mathrm{e}^{j\omega})\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|}{\sqrt{\alpha}}\right)d\omega,\ \zeta\triangleq 1-\frac{a}{b}$, and

$G(\mathrm{e}^{j\omega}) \triangleq \max\left\{\frac{a}{b}\frac{\sqrt{\alpha}}{2}\ ,\ \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|\right\}, \quad \forall\omega\in[-\pi,\pi]$

**Table 3.2:** Optimal Power Spectral Densities. $S_{\rm w}(\mathrm{e}^{j\omega})$ is the PSD of the output of the scalar quantizer; $S_\epsilon(\mathrm{e}^{j\omega})$ is the PSD of the frequency weighted reconstruction error, see Fig. 3.1.

| Given | Optimal $S_{\rm w}(\mathrm{e}^{j\omega})$ | Optimal $S_\epsilon(\mathrm{e}^{j\omega})$ | Where |
|---|---|---|---|
| $A$, $f$ | $\lvert A\Omega_{\rm x}\rvert^2+\sigma_{\rm n}^2 f^2$ | $\frac{\sigma_{\rm n}^2\lvert\Omega_{\rm x}P\rvert^2 f^2\left(a^2\sigma_{\rm n}^2+b^2\lvert\Omega_{\rm x}A\rvert^2\right)}{\left(a\sigma_{\rm n}^2+b\lvert\Omega_{\rm x}A\rvert^2\right)^2}$ | - |
| $B$, $f$ | $\frac{b^2\lvert P\rvert^4\lvert\Omega_{\rm x}B\rvert^2}{\left(aV+b\lvert PB\rvert^2\right)^2}+\sigma_{\rm n}^2 f^2$ | $\frac{a^2V^2\lvert\Omega_{\rm x}P\rvert^2}{\left(aV+b\lvert PB\rvert^2\right)^2}+\sigma_{\rm n}^2\lvert PB\rvert^2 f^2$ | (i) |
| $f$, $AB=W$ | $\kappa^2\lvert\Omega_{\rm x}P\rvert\lvert W\rvert f+\sigma_{\rm n}^2 f^2$ | $\lvert 1-W\rvert^2\lvert\Omega_{\rm x}P\rvert^2+\frac{\sigma_{\rm n}^2\lvert\Omega_{\rm x}W\rvert f}{\kappa^2\lvert P\rvert}$ | - |
| $f$ | $\kappa^2 Gf+\left(\sigma_{\rm n}^2-\kappa^2\xi\right)f^2$ | $\frac{\sigma_{\rm n}^2}{\kappa^2}Gf+\left(\xi-\frac{\sigma_{\rm n}^2}{\kappa^2}\right)\xi f^2$ | (ii) |
| $A$, $B$ | $\lvert\Omega_{\rm x}A\rvert^2+\sigma_{\rm n}^2\frac{K\lambda}{\lambda+\lvert PB\rvert^2}$ | $\lvert 1-AB\rvert^2\lvert\Omega_{\rm x}P\rvert^2+\sigma_{\rm n}^2\frac{K\lambda\lvert PB\rvert^2}{\lambda+\lvert PB\rvert^2}$ | (iii) |
| $B$ | $\frac{b^2\lvert P\rvert^4\lvert\Omega_{\rm x}B\rvert^2}{\left(a\lambda+b\lvert PB\rvert^2\right)^2}+\sigma_{\rm n}^2\frac{K\lambda}{\lambda+\lvert PB\rvert^2}$ | $\frac{a^2\lambda^2\lvert\Omega_{\rm x}P\rvert^2}{\left(a\lambda+b\lvert PB\rvert^2\right)^2}+\sigma_{\rm n}^2\frac{K\lambda\lvert PB\rvert^2}{\lambda+\lvert PB\rvert^2}$ | (iv) |
| $H$ | $\frac{b^2\lvert\Omega_{\rm x}\rvert^4\lvert PH\rvert^2}{\left(a\lambda+b\lvert\Omega_{\rm x}H\rvert^2\right)^2}+\sigma_{\rm n}^2\frac{K\lambda}{\lambda+\lvert\Omega_{\rm x}H\rvert^2}$ | $\frac{a^2\lambda^2\lvert\Omega_{\rm x}P\rvert^2}{\left(a\lambda+b\lvert\Omega_{\rm x}H\rvert^2\right)^2}+\sigma_{\rm n}^2\frac{K\lambda\lvert\Omega_{\rm x}H\rvert^2}{\lambda+\lvert\Omega_{\rm x}H\rvert^2}$ | (v) |
| - | $\kappa^2\frac{\sqrt{K\alpha}}{2}$ | $\frac{\alpha}{4}\left(1-\frac{\alpha\left[1-\frac{a^2}{b^2}\right]}{\left(\sqrt{G^2+\left[1-\frac{a}{b}\right]\alpha}+G\right)^2}\right)$ | (vi) |

(i): $V\triangleq\frac{\|PBf\|^2}{K-\|f\|^2}$

(ii): $\kappa>0$ is an arbitrary real constant, $\xi$ is the unique scalar that satisfies
$K=\frac{a}{b}\frac{1}{2\pi}\int_{-\pi}^{\pi}\max\left\{f(\mathrm{e}^{j\omega})^2\ ,\ \frac{f(\mathrm{e}^{j\omega})\left|\Omega_{\rm x}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|}{\xi}\right\}d\omega+[1-\frac{a}{b}]\|f\|^2$, and
$G(\mathrm{e}^{j\omega})\triangleq\max\left\{\xi f(\mathrm{e}^{j\omega})\ ,\ \left|\Omega_{\rm x}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|\right\},\quad\forall\omega\in[-\pi,\pi]$

(iii): $\ln(K)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left(\frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{\lambda}+1\right)d\omega.$

(iv): $\ln(K)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left(\frac{\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2}{\lambda}+1\right)d\omega.$

(v): $\ln(K)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left(\frac{\lambda_2}{\lambda_2-\lvert\Omega_{\rm x}(\mathrm{e}^{j\omega})A(\mathrm{e}^{j\omega})\rvert^2}\right)d\omega$

(vi): $\frac{1}{2}\ln(K)=\frac{1}{2\pi}\int\limits_{|\Omega_{\rm x}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})|\geq\frac{a\sqrt{\alpha}}{2b}}\ln\left(\sqrt{\frac{\left|P(\mathrm{e}^{j\omega})\Omega_{\rm x}(\mathrm{e}^{j\omega})\right|^2}{\alpha}+\zeta}\ +\ \frac{\left|P(\mathrm{e}^{j\omega})\Omega_{\rm x}(\mathrm{e}^{j\omega})\right|}{\sqrt{\alpha}}\right)d\omega,\ \zeta\triangleq1-\frac{a}{b}$, and
$G(\mathrm{e}^{j\omega})\triangleq\max\left\{\frac{a}{b}\frac{\sqrt{\alpha}}{2}\ ,\ \left|\Omega_{\rm x}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|\right\},\quad\forall\omega\in[-\pi,\pi]$

error, $S_\epsilon(\mathrm{e}^{j\omega})$, for each scenario. Careful analysis of these expressions reveals that, except for specific cases of $\Omega_\mathrm{x}$, $P$ and the given frequency responses for each scenario, $S_\epsilon(\mathrm{e}^{j\omega})$ is not constant over $[-\pi, \pi]$. However, when all three degrees of freedom are available for design (last row in Table 3.2), and if $a = b$, then $S_\epsilon(\mathrm{e}^{j\omega})$ is constant for all $\omega$ such that $\left|P(\mathrm{e}^{j\omega})\right|^2 S_\mathrm{x}(\mathrm{e}^{j\omega}) \geq (\frac{a}{b})^2 \frac{\alpha}{4} = \frac{\alpha}{4}$.

### 3.10.3 Optimal Performance

The equations characterizing the optimal trade-off between quantizer SNR and WCMSE are listed in Table 3.3. This table is a summary of the rate-distortion (or more precisely, SNR-distortion) results derived in Theorems 3.1–3.10. Using these results, it is possible to determine the SNR-distortion effect of having any subset of the filters $A(z)$, $B(z)$ and $F(z)$ fixed and given. Only in the scenarios corresponding to the first three rows of Table 3.3 is it possible to find $D_{a,b}$ directly from a given value of $K$. In all the other scenarios, $D_{a,b}(\mathrm{x}, \mathrm{y})$ is a bijective function of $K$. In some cases, $D_{a,b}(\mathrm{x}, \mathrm{y})$ and $K$ are connected by a scalar parameter, which needs to be determined numerically. The latter is not a big difficulty since, in all cases $K$ is related to these scalar parameters ($\alpha$ or $\lambda$) through monotonic functions. Moreover, if $a \leq b$, then it can be shown that the functions that relate $K$ to these scalar parameters are convex.

## 3.11 Simulation Example

To illustrate our results, we present below an example in which we design the filters of a Perfect Reconstruction FQ aimed at digitally encoding audio signals in a psycho-acoustically optimal manner. Recall that an optimal Perfect Reconstruction FQ is obtained by solving Optimization Problem 3.8 with $a = 1$ and $b \to \infty$, see also Section 3.9.1.

The details of the simulation model, as well as the results of both the simulations and the numerical optimizations are given below.

### 3.11.1 Simulation Setup

The PSD of audio signals was modeled as unit-variance zero mean white Gaussian noise filtered through $\Omega_\mathrm{x}(z) = 0.09315 \left(\frac{z+0.6773}{z-0.8588}\right)$. The magnitude of the frequency response of $\Omega_\mathrm{x}(z)$ is depicted in Fig. 3.3 (solid line). The frequency weighting filter $P(z)$ considered has a frequency response magnitude which approximates the psycho-acoustic curve derived in [46, Table 1], thus modelling the sensitivity of human hearing to noise[10]. The corresponding frequency response is plotted with a dotted line in Fig. 3.3 (the sampling frequency is $44.1$ [kHz]). The resulting $g = |\Omega_\mathrm{x} P|$ for these $\Omega_\mathrm{x}$ and $P(z)$ is also shown in the

[10]The coefficients of $P(z)$ can be found at http://msderpich.no-ip.org/research

**Table 3.3:** Minimum $D_{a,b}(\mathrm{x},\mathrm{y})$ and $K=\gamma+1$ ($\gamma$ is the SNR of $\mathcal{Q}$)

| Given | Minimum $D_{a,b}(\mathrm{x},\mathrm{y})$ | $K=\gamma+1$ |
|---|---|---|
| $A$, $f$ | $ab\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\lvert\Omega_{\mathrm{x}}P\rvert^2 f^2}{af^2+b\frac{\lvert\Omega_{\mathrm{x}}A\rvert^2}{\sigma_{\mathrm{n}}^2}}d\omega$ | $K=\frac{\lVert\Omega_{\mathrm{x}}A\rVert^2}{\sigma_{\mathrm{n}}^2}+\lVert f\rVert^2$ |
| $B$, $f$ | $ab\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\lvert\Omega_{\mathrm{x}}P\rvert^2}{a+b\frac{\lvert PB\rvert^2}{V}}d\omega$ | $K=\frac{\lVert PBf\rVert^2}{V}+\lVert f\rVert^2$ |
| $f$, $AB=W$ | $a\frac{\langle f,\lvert\Omega_{\mathrm{x}}P\rvert\lvert W\rvert\rangle^2}{K-\lVert f\rVert^2}+b\lVert(W-1)\Omega_{\mathrm{x}}P\rVert^2$ | $K$ |
| $f$ | $b\xi\langle f,\lvert\Omega_{\mathrm{x}}P\rvert\rangle_{\xi f<\lvert\Omega_{\mathrm{x}}P\rvert}+b\lVert\Omega_{\mathrm{x}}P\rVert^2_{\xi f>\lvert\Omega_{\mathrm{x}}P\rvert}$ | $K=\frac{a}{b}\frac{1}{2\pi}\int_{-\pi}^{\pi}\max\left\{f^2\ ,\ \frac{f\lvert\Omega_{\mathrm{x}}P\rvert}{\xi}\right\}d\omega+\left[1-\frac{a}{b}\right]\lVert f\rVert^2$ |
| $A$, $B$ | $a\lambda\lVert\Omega_{\mathrm{x}}A\rVert^2+b\lVert(AB-1)\Omega_{\mathrm{x}}P\rVert^2$ | $K=\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left(\frac{\lvert PB\rvert^2}{\lambda}+1\right)d\omega\right)$ |
| $B$ | $ab\lambda\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\lvert\Omega_{\mathrm{x}}P\rvert^2}{a\lambda+b\lvert PB\rvert^2}d\omega$ | $K=\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left(\frac{\lvert PB\rvert^2}{\lambda}+1\right)d\omega\right)$ |
| $H$ | $ab\lambda\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\lvert\Omega_{\mathrm{x}}P\rvert^2}{a\lambda+b\lvert\Omega_{\mathrm{x}}H\rvert^2}d\omega$ | $K=\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left(\frac{\lvert\Omega_{\mathrm{x}}H\rvert^2}{\lambda}+1\right)d\omega\right)$ |
| - | $\frac{a}{2\pi}\int_{\lvert\Omega_{\mathrm{x}}P\rvert\geq\frac{a\sqrt{\alpha}}{2b}}\frac{(\alpha/2)\lvert\Omega_{\mathrm{x}}P\rvert}{\sqrt{\lvert\Omega_{\mathrm{x}}P\rvert^2+\zeta\alpha}+\lvert\Omega_{\mathrm{x}}P\rvert}d\omega+\frac{b}{2\pi}\int_{\lvert\Omega_{\mathrm{x}}P\rvert<\frac{a\sqrt{\alpha}}{2b}}\lvert\Omega_{\mathrm{x}}P\rvert^2 d\omega$ | $K=\exp\left(\frac{1}{\pi}\int_{\lvert\Omega_{\mathrm{x}}P\rvert\geq\frac{a\sqrt{\alpha}}{2b}}\ln\left[\sqrt{\frac{\lvert P\Omega_{\mathrm{x}}\rvert^2}{\alpha}+\zeta}+\frac{\lvert P\Omega_{\mathrm{x}}\rvert}{\sqrt{\alpha}}\right]d\omega\right)$ |
| where $\zeta\triangleq 1-\frac{a}{b}$. | | |

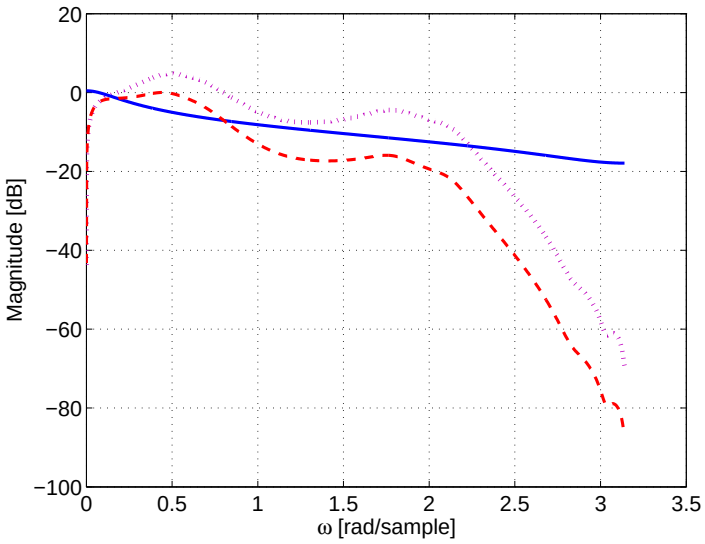



**Figure 3.3:** Frequency response magnitudes for $\Omega_{\mathrm{x}}(z)$ (solid line), $P(z)$ (dotted line) and $g(\omega) = \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$ (dashed line). The underlying sampling frequency is $44.1$ [KHz].

same figure (dashed line). For this choice of $g$, and in view of (3.149), one could expect the norm of a full whitening feedback filter to be very large. This is indeed the case: $\|F_0\|^2 = 2.2 \times 10^{11}$. Thus, the sub-optimal feedback filter characterized by (3.145) requires the use of a scalar quantizer with at least $18$ bits in order to become feasible (see Constraint 3.2).

In the simulations, $\mathcal{Q}$ was chosen to be a uniform mid-rise quantizer with quantization interval $\Delta = 1$. Several values of $\gamma$ were considered for the simulations, calculated as $\gamma = \frac{3}{\rho^2}2^{2b}$, where $b \in \{1, 2, \ldots, 16\}$ and where $\rho \triangleq \frac{N\Delta}{2\sigma_v}$ denotes the *loading factor*. Two different loading factors were considered: $4$ and $6$. The latter choice yields a slightly lower $\gamma$ than the usual loading factor of $4$. However, this regime has the benefit of making overload errors smaller and more infrequent. As the simulation results will show, for our choices of $\Omega_{\mathrm{x}}$ and $P$, this more conservative loading factor yields lower overall distortion when $b$ takes values above $6$ bits per sample.

For each $b$ (and corresponding two values for $\gamma$, one for each loading factor), the filters of the converter were designed according to the following:

1. The parameter $\alpha_{opt}$ was calculated by numerically solving (3.142).

2. The optimal $|1 - F|$, $|A|$ and $|B|$ were obtained via (3.140) and (3.22).

3. These functions were then approximated[11] with rational IIR transfer functions $A(z)$, $B(z)$ (of order 7) and $F(z)$ (of order 15).

4. An appropriate value for the parameter $\kappa$ in (3.41) was chosen via $\kappa^2 = 2\sigma_n^2\sqrt{\frac{K}{\alpha}}$, see (3.120), assuming $\sigma_n^2 = 1/12$ (recall that $\Delta = 1$ for all the simulations). This ensures that $\sigma_v^2 = \gamma\sigma_n^2$.

For each combination of $b$ and $\rho$, the resulting PRFQ converter was simulated utilizing two different architectures.

1. *Non Overloading $\mathcal{Q}$:* This scheme is as depicted in Fig. 3.1, with $\mathcal{Q}$ having (virtually) infinitely many levels. Thus, $|n(k)| \leq \frac{\Delta}{2}$ for all $k$ (neither clipping nor overload errors occur).

2. *Overloading $\mathcal{Q}$ and Clipped $n$:* Here, $\mathcal{Q}$ has $N = 2^b$ levels, which yields a scalar quantizer with a finite input dynamic range $[-N\frac{\Delta}{2}, N\frac{\Delta}{2}]$. As a consequence, any value $|v(k)| > N\frac{\Delta}{2}$ would overload $\mathcal{Q}$ (if $s = \infty$) or produce clipping error (if $s = V$). To avoid large limit-cycle oscillations, this variant was simulated using clipping (i.e., $s = V$).

Each simulation with the non-overloading PRFQ comprised $10^5$ samples. For the overloading converter, five $10^5$ samples simulations were performed for each combination of $\rho$ and $b$.

### 3.11.2 Results

The results of the numerical optimization and the simulations are discussed next.

#### Comparison between $D^\star$ and the Rate-Distortion Function

The information theoretic lower bound (see [148]) for the *frequency weighted* MSE (FWMSE) associated with the given source $\{x(k)\}_{k\in\mathbb{Z}}$ and filter $P(z)$ is plotted in Fig. 3.4 (solid line). This corresponds to Shannon's quadratic Distortion-Rate function $D(R)$ when $R = b$. As the bit-rate is increased, the gap between $D^\star$ and this absolute lower bound decreases to approx 7.5 [dB] for $\rho = 4$ and 11 [dB] for $\rho = 6$, at $b = 16$. This difference can be attributed to the rate-distortion inefficiency of the uniform scalar quantizer[12]. On the other hand, the larger performance gap observed at lower bit-rates can be attributed to the perfect reconstruction constraint.[13] Recall that, at low bit rates, the achievement of

[11] The optimization routines utilized are based upon the Matlab optimization toolbox and can be found at http://msderpich.no-ip.org/research.

[12] From Shannon's Rate-Distortion function for memoryless Gaussian sources, the maximum SNR for a bit-rate $b$ is $2^{2b}$. The SNR (neglecting overload errors) for a uniform scalar quantizer with loading factor $\rho$ is given by $\frac{3}{\rho^2}2^{2b}$. Thus, the theoretical performance gaps for $\rho = 4$ and 6 are $10\log_{10}(3/16) = 7.3$ [dB] and $10\log_{10}(3/36) = 10.8$ [dB], respectively.

[13] The quadratic Gaussian rate-distortion function with the constraint that the end-to-end distortion is uncorrelated to the source has recently been characterized by the author in [127]. The latter is also described in Section 4.5.2 of this thesis.

Shannon's rate-distortion function demands the suppression of relatively less significant bands of the PSD of the input signal (see, e.g., [6], and [148]). This linear distortion, which a PRFQ cannot achieve, is more severe at lower bit-rates. Thus, the performance gap increases as $b$ is reduced.

**Non Overloading $\mathcal{Q}$**

The FWCMSE of this form of converter is presented in four of the plots in Fig. 3.4, with labels beginning with "$\sigma_\epsilon^2$ opt. PRFQ, Non Overloading". These plots differ in the loading factor, denoted by "$\rho$", and in the meaning of $b$ in each case. For the plots whose labels do not have the ending "E.C." (entropy coding), $b$ is simply the number utilized to generate the value $\gamma = \frac{3}{\rho^2}2^{2b}$ for which the filters were optimized. The plots whose labels end in "E.C." correspond to the same simulations, but for each point the value of $b$ is the *scalar* entropy of the quantized output of the converter. It can be seen in Fig. 3.4 that the FWCMSE obtained for the non overloading $\mathcal{Q}$ without entropy coding is remarkably close to the theoretical value $D^\star$ predicted by (3.143). More importantly, even for bit-rates as small as $b = 2$, each observed ratio $\sigma_v^2/\sigma_n^2$ deviates from its nominal value of $\gamma$ by less than $2\%$. (For the extreme situation $b = 1$, the observed $\sigma_v^2$ was slightly lower than predicted, while $\sigma_n^2$ was $55\%$ higher than $1/12$ due to the highly non-uniform PDF of the resulting sequence $\{n(k)\}_{k\in\mathbb{Z}}$.) It can also be seen that the scalar entropy of the quantized output of the PRFQ in these cases is very close to Shannon's $R(D)$ function for a given distortion. This agrees with the observation that the output of $\mathcal{Q}$ in an optimized PRFQ is white, see Section 3.10.2. The difference between these quantities is bigger for lower values of $b$, for the same reason given in the previous paragraph.

**Overloading $\mathcal{Q}$**

For the Overloading PRFQ using an $\rho$ of $4$, the FWCMSE diminishes along with the corresponding $D^\star$ for $b \in \{1, \ldots, 6\}$. However, the measured FWCMSE varies very little for $b \geq 7$, staying several dB higher than $D^\star$ over that range of bit-rates. This performance degradation can be attributed to clipping errors. The fact that overload errors become noticeable only for high bit rates (many quantization levels) might seem, at first, surprising. However, this phenomenon can be easily explained by noting that the size of the tails of the PDF of $\{v(k)\}_{k\in\mathbb{Z}}$ that fall outside the dynamic range of $\mathcal{Q}$ remains approximately constant in relation to $N\Delta = 2^b\Delta$ for all $b$. (This is a direct consequence of the loading factor rule.) In contrast, granular (non-overloading) quantization error is proportional to $\Delta^2$, which is held constant in the simulations. Therefore, the ratio between clipping and granular quantization errors grows approximately as $2^b$ and clipping errors become dominant for sufficiently high bit-rates.

Because of the reduced occurrence (and magnitude) of clipping errors, the optimized PRFQ with

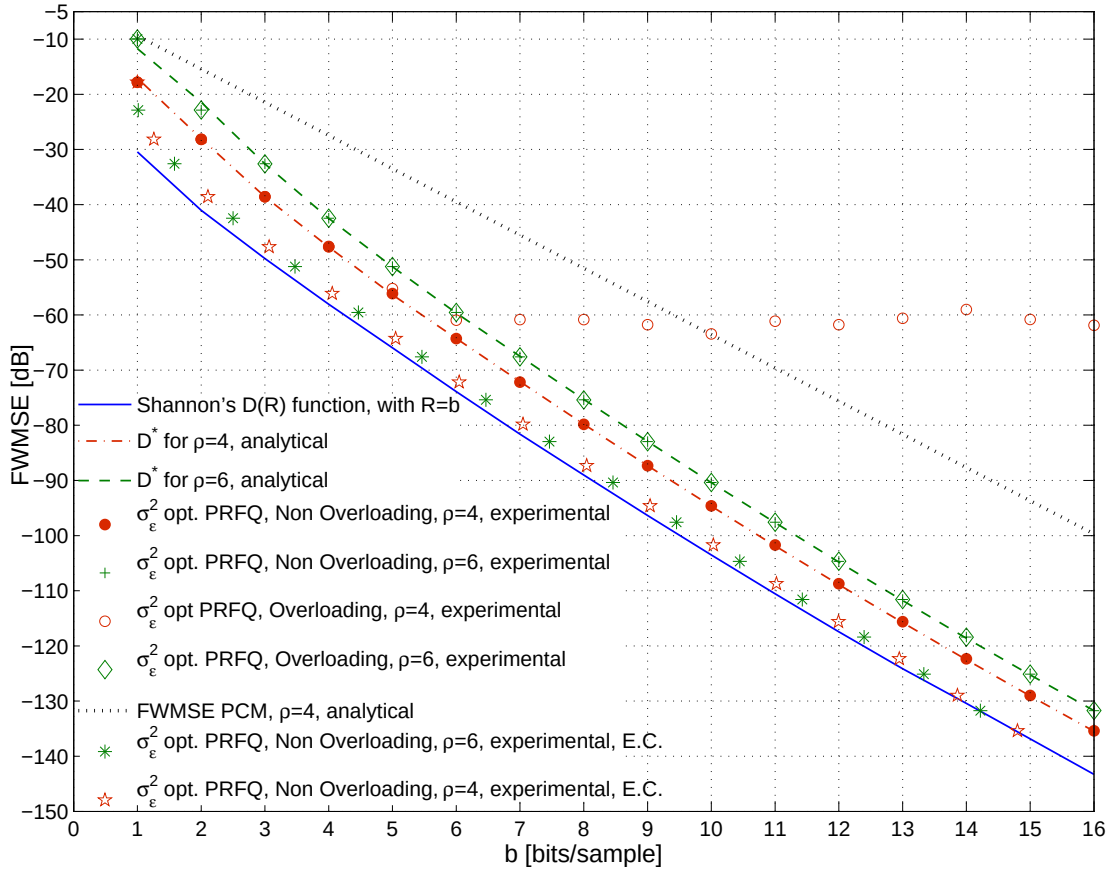



**Figure 3.4:** Frequency weighted MSE for $b \in \{1, \ldots, 16\}$.

overloading $\mathcal{Q}$ and $\rho = 6$ exhibits an FWCMSE smaller than that of its counterpart with $\rho = 4$ for $b \geq 7$. Furthermore, this more conservative loading factor allows the converter to perform almost exactly as predicted by our analytical expression for $D^\star$.[14]

### Comparison with PCM

The theoretical FWCMSE of a PCM A/D converter, denoted by $D_{PCM}$, can be found from (3.23) by taking $a = 1$ and making $A(z) \equiv B(z) \equiv 1$ and $F \equiv 0$, which gives $D_{PCM} = \|\Omega_{\mathrm{x}}\|^2 \|P\|^2/\gamma$. For the chosen input PSD and frequency weighting filter, and calculating $\gamma$ as $\frac{3}{16}2^{2b}$, the value of $D_{PCM}$ varies with $b$ as shown in Fig. 3.4 (dotted line). As seen in this figure, the gap between $D^\star$ and $D_{PCM}$, for each

[14] There exist several results on the optimal balance between overload and granular error variances for stand-alone scalar quantizers (see, e.g., [94] and the references therein). However, for feedback quantizers the question seems to be open. An optimal trade-off between overload and granular errors as the oversampling ratio tends to infinity is found in Theorem 3.15, see Section 3.12.3.

value of $\rho$, gets smaller as the bit-rate decreases. It can also be seen in Fig 3.4 that the optimized PRFQ with overloading and $\rho = 6$ exhibits an improvement of $32$ [dB] over PCM at $b = 16$. Equivalently, in order to obtain the same FWCMSE as that of PCM at $16$ bits, the PRFQ converter with $\rho = 6$ requires less than $12$ bits. At lower bit-rates, the improvement of the optimal PRFQ over PCM is also significant. For example, the overloading PRFQ with $\rho = 4$ and $b = 2$ has a lower FWCMSE than the PCM converter with $b = 4$, thus achieving a data rate compression of $50\%$ (see Fig. 3.4).

## 3.12 Oversampled Feedback Quantization

### 3.12.1 Introduction

As already mentioned in Section 1.1.4, there exist situations in which increasing the rate at which a continuous time source is sampled is preferable (or less expensive) than improving the accuracy of the quantization by increasing the number of quantization levels [66], [43, Section 1.1]. The use of oversampling along with scalar quantization is known to reduce reconstruction MSE for a given number of quantization levels. Using the Linear Model (defined in Sec 3.2.2), it has been shown in [56] that the MSE of scalar feedback quantizers can be made to decay with the oversampling ratio $\lambda$ as

$$MSE = \mathcal{O}(\lambda^{-2(m+1)}), \quad \text{when } \lambda \to \infty, \tag{3.150}$$

where $m$ is the order of the feedback filter (see also recent work in [92]). Of course, if the number of quantization levels in the scalar quantizer is kept fixed, and if memoryless entropy coding (or no entropy coding at all) is utilized, then the operational bit-rate increases proportionally with $\lambda$. Thus, the decay rate in (3.150) is rate-distortion inefficient, since a linear increase in the bit-rate at a fixed sampling ratio reduces MSE as $\mathcal{O}(2^{-2b\lambda})$, i.e., exponentially.

Recent work in [96] has shown that, for sources with bounded support, $1$-bit $\Sigma\Delta$ quantization can attain an MSE which decays as

$$MSE = \mathcal{O}(2^{-0.14\lambda}), \quad \text{when } \lambda \to \infty. \tag{3.151}$$

Such exponential decay rate is obtained by selecting a different feedback filter for each oversampling ratio [96]. This result was obtained using a deterministic model of quantization errors, and, to the best of the author's knowledge, corresponds to the fastest decay ratio of the MSE with oversampling available in the literature. Unfortunately, applying the method utilized in [96] for the cases in which the source has unbounded support, or to multi-bit feedback quantizers, seems to be a formidable task.

In Section 3.12.2 we will show that, within the Linear Model, if the optimal infinite order filters characterized in Section 3.9 are used for each value of $\lambda$, then one can achieve an exponential decay of

$D^\star$ with oversampling ratio, provided $\gamma$ is kept constant. For simplicity, we will restrict the analysis to the case in which the weights of the WCMSE are $a = 1$, $b \to \infty$, i.e., where the WCMSE reduces to the MSE, and the reconstruction error is uncorrelated to the source. We will then link $\gamma$ to the operational bit-rate of the scalar quantizer, and obtain asymptotic MSE decay rates when the operational bit-rate is kept constant. It will be shown below that, using an entropy coded subtractively dithered scalar quantizer (SDUSQ), the MSE of an optimal PR feedback quantization decreases as

$$MSE = \mathcal{O}(2^{-1.746\lambda}), \quad \text{when } \lambda \to \infty, \tag{3.152}$$

when the operational bit-rate is kept fixed, provided sufficient quantization levels to avoid clipping/overload are employed.

This result is then extended, in Section 3.12.3, to the cases in which the FQ uses clipping and a subtractively dithered scalar quantizer with $N$ levels. With this setting, we will demonstrate that, by adjusting the loading factor $\rho$ to each oversampling ratio, the MSE can be made to decay as

$$MSE = \mathcal{O}(\mathrm{e}^{-c_0\lambda^{1/3}}), \quad \text{when } \lambda \to \infty, \tag{3.153}$$

where $c_0 \triangleq [0.5(N-1)]^{2/3}$. This asymptotic behaviour of the MSE holds for sources with bounded or unbounded support, provided condition (3.1) on page 45 is satisfied.

### 3.12.2 The Oversampled Case Without Clipping/Overload

If the input sequence $\{x(k)\}_{k\in\mathbb{Z}}$ is obtained by sampling a band-limited analog signal, oversampling would cause $g$ (defined in (3.141)) to vary with $\lambda$. To capture this effect, we replace $g$ by the family of functions $g_\lambda$, defined as

$$g_\lambda(\omega) \triangleq \begin{cases} \sqrt{\lambda}\, g_1(\lambda\omega) & \text{, if } \quad |\omega| < \omega_c, \\ 0 & \text{, if } \omega_c \le |\omega| \le \pi. \end{cases} \tag{3.154}$$

In (3.154), $g_1$ denotes the square root of the PSD of the frequency weighted input without oversampling, and $\omega_c \triangleq \frac{\pi}{\lambda}$. Notice that $\|g_\lambda\|^2$, that is, the total power of $g_\lambda$ (in units of variance per sample), remains constant for all $\lambda \ge 1$. This ensures a uniform comparison basis for the distortion figures.

When considering the asymptotic performance of oversampled quantization as $\lambda \to \infty$, the validity of Assumption 3.2 on page 49 needs to be reconsidered. To see this, notice that if the number of quantization levels is insufficient to avoid clipping/overload errors, and if dither and clipping are used with a fixed loading factor, then there always exists a certain finite value of $\lambda$ beyond which Assumption 3.2 is violated. This arises from the fact that, for any fixed loading factor, the effect of clipping errors in the output does not decay with $\lambda$, thus becoming the dominant component in the FWMSE for sufficiently

high oversampling ratios. Further reduction of the FWMSE would then require one to balance clipping and granular quantization errors by increasing the loading factor. If the number of quantization levels is fixed, this would necessarily reduce the value of $\gamma$, clearly increasing the component of the FWMSE due to granular quantization errors. Nevertheless, if clipping and dither are used (with $s = V$), then the Linear Model and Theorem 3.12 are exact in describing the FWMSE due to *granular* quantization errors. Furthermore, as explained in Remark 3.1 below, for subtractively dithered quantization with infinitely many quantization levels, the entropy of the quantized output conditioned to the dither can be kept constant (and finite) as $\lambda \to \infty$ while having the reconstruction MSE that decays exponentially.

We can now make explicit the dependence of $D^\star$ on $\gamma$ and $\lambda$ by writing

$$D^\star(K,\lambda) \triangleq \min_{\substack{f\in\mathcal{C}_2\cap\mathcal{C}_1\\ g=g_\lambda}} D(f) = \min_{f\in\mathcal{C}_2\cap\mathcal{C}_1} \frac{\langle f, g_\lambda\rangle^2}{K - \|f\|^2}, \tag{3.155}$$

where $K$, defined in (3.22), corresponds to the *output-SNR of* $\mathcal{Q}$.

Interestingly, it is possible to establish a precise "exchange" formula for $K$ and $\lambda$. Indeed, in terms of minimal achievable distortion, the effect of increasing oversampling is equivalent to an exponential increase in the output-SNR of $\mathcal{Q}$. This is shown in the next theorem:

**Theorem 3.11.** Under the Linear Model described in Section 3.2.2, for any function $g_1(\omega)$, and for any $K > 1$, $\lambda \geq 1$, the minimum achievable FWMSE satisfies:

$$D^\star(K,\lambda) = D^\star(K^\lambda, 1). \tag{3.156}$$

▲

If we assume that $\gamma$ depends exponentially on the number of bits per sample, then Theorem 3.11 suggests an FWMSE that decays exponentially with $\lambda$, provided the Linear Model holds and that optimal filters $A(z)$, $B(z)$ and $F(z)$ (characterized by (3.140) and (3.22)) are employed for each $\lambda$. The following simple example illustrates this idea:

***Example*** *(Flat Weighted Input Spectrum) Consider an input signal $\{x(k)\}_{k\in\mathbb{Z}}$ and a weighting filter $P(z)$ such that $|\Omega_{\mathrm{x}}P|$ is constant $\forall\omega \in [-\pi,\pi]$, without oversampling. For this setup, the optimal $F(z)$ for our model of PRFQ is $F(z) \equiv 0$ ($f(\mathrm{e}^{j\omega}) \equiv 1$), i.e., a PCM converter. From* (3.143)*, the minimum frequency weighted MSE without oversampling (i.e., with $\lambda = 1$) under these conditions becomes*

$$D^\star(K,1) = \frac{\sigma_{xP}^2}{\gamma} = \frac{\sigma_{xP}^2}{K-1},$$

*where $\sigma_{xp}^2 \triangleq \|\Omega_{\mathrm{x}}P\|^2$. To analyze oversampling behaviour of $D^\star$ in this case, we apply Theorem 3.11 to the above expression. This gives that $D^\star(K,\lambda) = \frac{\sigma_{xP}^2}{K^\lambda - 1}$, and, thus,*

$$\sigma_{xP}^2 K^{-\lambda} \leq D^\star(K,\lambda) \leq \left(\frac{\sigma_{xP}^2}{1-K^{-1}}\right) K^{-\lambda} \tag{3.157}$$

*for all* $\lambda \geq 1$. *Note that, to achieve* (3.157), $F(z)$ *needs to be synthesized according to* (3.140a), (3.142) *and* (3.22). *Therefore, for this example, the MSE of an optimized PRFQ with fixed* $\gamma$ *exhibits an exponential decay with the oversampling ratio (since, by definition,* $K > 1$*).*

*If we further assume* $K$ *to depend on the number of bits per sample* $b$ *as* $K = \frac{3}{16}2^{2b} + 1$ *(which would correspond to* $\mathcal{Q}$ *being a uniform quantizer with many levels and operating with a loading factor of* $4$*), then* (3.157) *becomes*

$$\begin{aligned} &\sigma_{xP}^2\, 2^{-\left\lceil \log_2(\frac{3}{16}+2^{-2b})+2b \right\rceil \lambda} \leq \\ &\qquad D^\star(K,\lambda) < \left(\tfrac{\sigma_{xP}^2}{1-K^{-1}}\right) 2^{-\left\lceil \log_2(\frac{3}{16}+2^{-2b})+2b \right\rceil \lambda}. \end{aligned} \tag{3.158}$$

*The term* $\log_2(\frac{3}{16}+2^{-2b})$ *in* (3.158) *is negative for all* $b \geq 1$. *This implies that the decrease of* $D^\star$ *with* $\lambda$*, although exponential, is slower than* $2^{-2b\lambda}$. *Thus, the use of oversampling in this case is rate-distortion inefficient. In particular, taking* $b = 1$*, and supposing that Assumptions 3.3 and 3.4 hold, we obtain from* (3.158) *that* $D^\star(K,\lambda)$ *is lower and upper bounded by terms proportional to* $2^{-0.807\lambda}$. *For loading factor values of* $6$, $10$ *and* $20$*, the exponent in the latter expression changes to* $-0.41\lambda$, $-0.1635\lambda$ *and* $-0.0426\lambda$*, respectively.* ▲

The next theorem shows that the exponential decay of the FWMSE obtained in the example above can be extended to arbitrary (band-limited) input signals and frequency weighting criteria.

**Theorem 3.12.** For any $K > 1$ and function $g_1(\omega)$ satisfying Assumption 3.1, the following holds:

$$D^\star(K,\lambda) \leq \frac{K^2 \alpha_{opt}(K,1)}{4(K-1)} K^{-\lambda}, \quad \forall K > 1, \forall \lambda \geq 1, \tag{3.159}$$

*where* $\alpha_{opt}(K,1)$ *denotes the optimal* $\alpha$ *for* $\lambda = 1$. ▲

Thus, under the Linear Model, we have that the FWMSE of an optimized PRFQ decays exponentially with $\lambda$.

**Remark 3.1.** *As already mentioned in Section 3.2.2, if* $x$ *is bounded and a sufficiently large number of quantization levels to avoid overload is used together with dither, then the Linear Model is exact. Nevertheless, there is no guarantee that the number of quantization levels necessary to avoid overload remains constant as* $\lambda$ *increases. Thus, if that number increases with* $\lambda$*, then keeping* $\gamma$ *constant may require increasing the number of quantization levels in the quantizer. Nevertheless, if a subtractively dithered scalar quantizer is used, then the entropy of its quantized output conditioned to the dither, denoted by* $R_Q$*, depends only on the PDF of* $\mathrm{v}(k)$ *and on the quantization interval* $\Delta$*, see Section 2.4.2.* $R_Q$ *corresponds to the asymptotically achievable rate that can be obtained by entropy coding long sequences of quantized values. Thus, when using SDUSQ with sufficiently (possibly infinitely) many levels and* fixed rate*, one can substitute* (2.60) *into* (3.159)*, which yields* $D^\star(K,\lambda) = \mathcal{O}(2^{-1.746\lambda})$. ▲

An extension of these results to include the effect of clipping errors, which are unavoidable if the source has unbounded support and the quantizer has a finite number of quantization levels, is the main result of the next section.

### 3.12.3 The Oversampled Case With Clipping

In this section we derive an upper bound on the total frequency-weighted MSE of a perfect reconstruction FQ, including clipping errors. To do so, we assume that the scalar quantizer in Fig. 3.1 uses subtractive dither, uniformly distributed over $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$. Choosing the saturation threshold of the clipper as $s = N\frac{\Delta}{2}$, no overload occurs (see Section 3.2.1), and thus

$$\sigma_{\mathrm{n}}^2 = \frac{\Delta^2}{12} \tag{3.160}$$

Denote the *loading factor* at which $\mathcal{Q}$ operates by

$$\rho \triangleq \frac{\sigma_{\mathrm{v}}}{N\Delta/2}. \tag{3.161}$$

Substitution of (3.160) and (3.161) into (3.15) yields

$$\gamma = \frac{\sigma_{\mathrm{v}}^2}{\sigma_{\mathrm{n}}^2} = \frac{N^2\Delta^2}{4\rho^2\Delta^2/12} = \frac{3}{\rho^2}N^2 \tag{3.162}$$

In order to keep clipping errors infrequent and small, it is required to choose $\rho$ large enough.

In the FQ shown in Fig. 3.1, clipping errors are not injected into the feedback loop. Instead, they appear in the frequency weighted error $\epsilon$ filtered by $B(z)$ and $P(z)$, as the process

$$\tilde{\vartheta}(k) \triangleq P(z)B(z)\vartheta(k), \quad \forall k \in \mathbb{Z}. \tag{3.163}$$

Unless the source $\{\mathrm{x}(k)\}$ is a stationary process, one cannot guarantee that the samples of the clipping error will form a stationary, or even a w.s.s. random process. In order to quantify its contribution to the FWMSE for non-necessarily stationary sources, we define the *average frequency weighted power of clipping errors in the output* as

$$\sigma_{\tilde{\varepsilon}_\vartheta}^2 \triangleq \lim_{\ell\to\infty} \frac{1}{2\ell+1} \sum_{k=-\ell}^{\ell} \mathrm{E}\left[\tilde{\vartheta}(k)^2\right] \tag{3.164}$$

The next lemma provides a fundamental result that will serve to derive an upper bound for clipping errors.

**Lemma 3.13.** *Let $\varsigma_1, \varsigma_2, \ldots$ be independent random variables with moments $\mu_n^i \triangleq E[\varsigma_i^n]$, and let $\sigma^2 \triangleq \sum_i \mu_2^i < \infty$. If there exists a constant $H$ such that*

$$\left|\mu_n^i\right| \leq \frac{1}{2}(n!)H^{n-2}\left|\mu_2^i\right|, \quad \forall n \geq 2, \forall i \in \mathbb{Z}^+, \tag{3.165}$$

*then*

$$Pr\left\{\sum\nolimits_i \varsigma_i > u\sigma\right\} \leq \mathrm{e}^{-\frac{\sigma}{2H}u}, \quad \forall u \geq \sigma/(2H). \tag{3.166}$$

▲

*Proof.* From one of Bernstein's inequalities, given in [149, Sec. 5.5], we have that

$$Pr\left\{\sum\nolimits_i \varsigma_i > u\,\sigma\right\} \leq \mathrm{e}^{-2(1-c)u^2}, \quad \forall u > 0 \text{ and } \forall c \in (0,1) \text{ such that } u \leq \frac{c\sigma}{2(1-c)H}. \tag{3.167}$$

For every $u > 0$, the tightest bound for the first inequality in (3.167) is obtained with $c = \frac{u}{\frac{\sigma}{2H}+u}$. Substituting this into (3.167) yields $Pr\{\sum_i \varsigma_i > u\sigma\} \leq \mathrm{e}^{-2u^2/(1+\frac{2H}{\sigma}u)}$. The latter, together with the fact that $2u/(1+\frac{2H}{\sigma}u) \geq \frac{\sigma}{2H}$, $\forall u \geq \sigma/(2H)$ leads directly to (3.166), completing the proof. □

The following theorem provides an upper bound for $\sigma^2_{\varepsilon_\vartheta}$ applicable to situations in which the source has unbounded support.

**Theorem 3.14.** *Suppose there exists a scalar $\hat{g} < \infty$ such that $g_1(\omega) \leq \hat{g},\ \forall \omega \in [-\pi, \pi]$, see* (3.154)*. Assume that the innovations of the process $\{\mathrm{x}(k)\}$ is a sequence of zero-mean, independent random variables $\{\xi(k)\}$ having a symmetric PDF and moments which satisfy* (3.165) *with $H = H_\xi$, for some constant $H_\xi$. Then, in an optimal PRFQ with clipping and subtractive dither,*

$$\sigma^2_{\varepsilon_\vartheta} \leq 16\frac{\hat{g}^2}{\nu^2}\lambda\,\mathrm{e}^{-\nu\rho}, \quad \forall \lambda \geq 1, \tag{3.168}$$

*where $\lambda$ denotes the oversampling ratio, $\rho$ is the loading factor defined in* (3.161)*, and where*

$$\nu \triangleq \frac{1}{2}\min\left\{\left(\frac{\gamma\lambda}{\gamma+1}\right)^{1/2}\frac{\sigma_\xi}{H_\xi}, \frac{\sigma_\mathrm{n}}{H_n}\right\}. \tag{3.169}$$

▲

*Proof.* We have from (3.164) that

$$\sigma^2_{\epsilon_\vartheta} \leq B^2_{max}\sigma^2_\vartheta, \tag{3.170}$$

where

$$B^2_{max} \triangleq \max_{\omega\in[-\pi,\pi]} \left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 \tag{3.171}$$

and where $\sigma^2_\vartheta$ is the time-averaged power of clipping errors, given by

$$\sigma^2_\vartheta \triangleq \lim_{\ell\to\infty}\frac{1}{2\ell+1}\sum_{k=-\ell}^{\ell}\mathrm{E}\left[\vartheta(k)^2\right]. \tag{3.172}$$

We will first upper bound $B_{max}$ and then $\sigma^2_\vartheta$.

**Bounding $B_{max}^2$** From (3.41b), we have:

$$\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 = \frac{g_\lambda(\omega)}{\kappa^2 f(\mathrm{e}^{j\omega})}, \quad \forall\omega \in [-\pi,\pi]. \tag{3.173}$$

From (3.120),

$$\kappa^2 = 2\sigma_{\mathrm{n}}^2\left(\frac{\gamma+1}{\alpha(K,\lambda)}\right)^{1/2}. \tag{3.174}$$

Substitution of the latter into (3.140a) yields

$$\kappa^2 f(\mathrm{e}^{j\omega}) = 2\sigma_{\mathrm{n}}^2\frac{\gamma+1}{\alpha(K,\lambda)}\left(\sqrt{g_\lambda(\omega)^2+\alpha(K,\lambda)}-g_\lambda(\omega)\right) = 2\sigma_{\mathrm{n}}^2\frac{\gamma+1}{\sqrt{g_\lambda(\omega)^2+\alpha(K,\lambda)}+g_\lambda(\omega)}.$$

Noting that $\alpha(K,\lambda) = \lambda\alpha(K^\lambda,1)$ (see the proof of Theorem 5 in [118]), we obtain

$$\begin{aligned}\kappa^2 f(\mathrm{e}^{j\omega}) &= 2\sigma_{\mathrm{n}}^2\frac{\gamma+1}{\sqrt{\lambda g_1(\lambda\omega)^2+\lambda\alpha(K^\lambda,1)}+\lambda g_1(\lambda\omega)} \\ &= \frac{2\sigma_{\mathrm{n}}^2(\gamma+1)}{\sqrt{\lambda}}\cdot\frac{1}{\sqrt{g_1(\lambda\omega)^2+\alpha(K^\lambda,1)}+g_1(\lambda\omega)}.\end{aligned}$$

Substituting this last equation and (3.154) into (3.173) we obtain

$$\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 = \frac{\lambda}{2\sigma_{\mathrm{n}}^2(\gamma+1)}\left(\sqrt{g_1(\lambda\omega)^2+\alpha(K^\lambda,1)}+g_1(\lambda\omega)\right)g_1(\lambda\omega), \quad \forall\omega\in[-\pi,\pi]. \tag{3.175}$$

Since $\alpha(K^\lambda,1)$ decreases monotonically with increasing $\lambda$, the following upper bound can be obtained from (3.175):

$$\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 \le \frac{\lambda}{2\sigma_{\mathrm{n}}^2(\gamma+1)}\left(\sqrt{\hat{g}^2+\alpha(K,1)}+\hat{g}\right)\hat{g}, \quad \forall\omega\in[-\pi,\pi]. \tag{3.176}$$

In order to get rid of $\alpha(K,1)$ in the above expression, we will use an upper bound for $\alpha(K,1)$ instead of the latter in the right hand side of (3.176). Since $K=\gamma+1$, it follows directly from (3.142) that

$$\alpha(K,1) \le 4\hat{g}^2\frac{\gamma+1}{\gamma^2}, \tag{3.177}$$

and thus

$$\left(\sqrt{\hat{g}^2+\alpha(K,1)}+\hat{g}\right)\hat{g} \le \left(\sqrt{1+4\frac{\gamma+1}{\gamma^2}}+1\right)\hat{g}^2 = 2\hat{g}^2\frac{\gamma+1}{\gamma}.$$

Substitution of the latter into (3.176) yields

$$\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 \le B_{max}^2 \le \frac{\hat{g}^2}{\sigma_{\mathrm{n}}^2\gamma}\lambda, \quad \forall\omega\in[-\pi,\pi]. \tag{3.178}$$

**Bounding $\sigma_\vartheta^2$** For every $k \in \mathbb{Z}$, $v(k)$ is a linear combination of the i.i.d. random variables $\{\mathrm{n}(i)\}_{i=-\infty}^{k-1}$, and the independent random variables $\{\xi(i)\}_{i=-\infty}^{k}$. Notice also that, due to the use of subtractive dither, the random variables $\mathrm{n}(i)$ and $\xi(k)$ are independent for all $i, k \in \mathbb{Z}$. More explicitly, at any instant $k$, we can write

$$\mathrm{r}(k) = \sum_{i=0}^{\infty} c_\xi(i)\xi(k-i); \qquad \mathrm{y}(k) = \sum_{i=1}^{\infty} c_n(i)\,\mathrm{n}(k-i); \qquad \mathrm{v}(k) = \sum_{i=0}^{\infty} \varsigma(i), \tag{3.179}$$

where the sequence

$$\varsigma(i) \triangleq \begin{cases} c_\xi(\frac{i}{2})\xi(k-\frac{i}{2}) & ,i \text{ even} \\ c_n(\frac{i+1}{2})\,\mathrm{n}(k-\frac{i+1}{2}) & ,i \text{ odd,} \end{cases} \tag{3.180}$$

is made of independent random variables. We will upper bound $\sigma_\vartheta^2$ by applying Lemma 3.13 to $\sum_{i=0}^{\infty}\varsigma(i)$. For this purpose we need to find a value for $H$, namely $H_\varsigma$, for which the random variables $\{\varsigma_i\}_{i=0}^{\infty}$ satisfy (3.165). This can be done by upper bounding the coefficients $c_\xi$ and $c_n$. From (3.179) and Fig. 3.1 we have

$$\sigma_\xi^2 \sum_{i=0}^{\infty} c_\xi(i)^2 = \sigma_{\mathrm{r}}^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|A(\mathrm{e}^{j\omega})\right|^2 g(\omega)^2 d\omega. \tag{3.181}$$

From (3.175), and since $A(\mathrm{e}^{j\omega}) = B(\mathrm{e}^{j\omega})^{\sim 1}$, the squared frequency response magnitude of the pre-filter $A(z)$ can be upper bounded as $\left|A(\mathrm{e}^{j\omega})\right|^2 \le \sigma_{\mathrm{n}}^2(\gamma+1)/(g(\omega)^2\lambda)$, which, when substituted into the right hand side of (3.181), yields $\sum_{i=-\infty}^{k} c_\xi(i)^2\sigma_\xi^2 \le \frac{\sigma_{\mathrm{n}}^2(\gamma+1)}{\lambda} = \sigma_{\mathrm{v}}^2\frac{\gamma+1}{\gamma\lambda}$. The latter immediately gives the upper bound

$$c_\xi(i)^2 \le \frac{\gamma+1}{\sigma_\xi^2\gamma\lambda}\sigma_{\mathrm{v}}^2, \quad \forall i \in \mathbb{Z}_0^+. \tag{3.182}$$

Similarly, from (3.179), and since $\sigma_{\mathrm{n}}^2 \le \sigma_{\mathrm{v}}^2$, we have that $\sum_{i=-\infty}^{k-1} c_n(i)^2\sigma_{\mathrm{n}}^2 \le \sigma_{\mathrm{v}}^2$, which leads directly to

$$c_n(i)^2 \le \frac{\sigma_{\mathrm{v}}^2}{\sigma_{\mathrm{n}}^2}, \quad \forall i \in \mathbb{Z}^+. \tag{3.183}$$

Since, for any random variable x and scalar $c$, $H_{c\,\mathrm{x}} = cH_{\mathrm{x}}$, it follows from (3.180), (3.182) and (3.183) that $H_\varsigma$ can be upper bounded as

$$H_\varsigma \le \hat{H}_\varsigma \triangleq \max\{\max_i\{c_\xi(i)\}H_\xi, \max_i\{c_n(i)\}H_n\} \le \max\left\{\left(\frac{\gamma+1}{\gamma\lambda}\right)^{1/2}\frac{H_\xi}{\sigma_\xi}, \frac{H_n}{\sigma_{\mathrm{n}}}\right\}\sigma_{\mathrm{v}} \tag{3.184}$$

Substituting $H$ by $H_\varsigma$ into (3.166) we obtain

$$Pr\{\mathrm{v} > u\sigma_{\mathrm{v}}\} \le \mathrm{e}^{-\frac{\sigma_{\mathrm{v}}}{2H_\varsigma}u} \le \mathrm{e}^{-\frac{\sigma_{\mathrm{v}}}{2\hat{H}_\varsigma}u}, \quad \forall u \ge \sigma_{\mathrm{v}}/(2H_\varsigma). \tag{3.185}$$

From (3.185) we have that the variance of $\vartheta$ cannot be larger than that obtained if $\mathrm{v}$ were a random variable with cumulative PDF given by

$$F_{\mathrm{v}_{max}}(x) \triangleq \begin{cases} \mathrm{e}^{\frac{1}{2\hat{H}_\varsigma}x} & \text{, if } x < -2\hat{H}_\varsigma \ln(2), \\ 1/2 & \text{, if } |x| \leq 2\hat{H}_\varsigma \ln(2), \\ 1-\mathrm{e}^{-\frac{1}{2\hat{H}_\varsigma}x} & \text{, if } x > 2\hat{H}_\varsigma \ln(2). \end{cases} \tag{3.186}$$

Hence, for $\rho > \frac{2\hat{H}_\varsigma}{\sigma_\mathrm{v}} \ln(2)$, the variance of overload errors can be upper bounded as

$$\sigma^2_\vartheta \leq \int_{\rho\sigma_\mathrm{v}}^{\infty} (t-\rho\sigma_\mathrm{v})^2 \Big[\frac{d}{dt} F_{v_{max}}(t)\Big] dt = \frac{2}{2\hat{H}_\varsigma} \int_{\rho\sigma_\mathrm{v}}^{\infty} (t^2 - 2\rho\sigma_\mathrm{v} t + \rho^2\sigma^2_\mathrm{v})\, \mathrm{e}^{-\frac{1}{2\hat{H}_\varsigma}t}\, dt \tag{3.187}$$

$$= 16\hat{H}^2_\varsigma\, \mathrm{e}^{-\frac{\sigma_\mathrm{v}}{2\hat{H}_\varsigma}\rho} = 16\hat{H}^2\, \mathrm{e}^{-\nu\rho}, \tag{3.188}$$

since $\frac{\sigma_\mathrm{v}}{2\hat{H}_\varsigma} = \frac{1}{2}\min\left\{\left(\frac{\gamma\lambda}{\gamma+1}\right)^{1/2}\frac{\sigma_\xi}{H_\xi}, \frac{\sigma_\mathrm{n}}{H_n}\right\} = \nu$, see (3.169). Substituting (3.188) and (3.178) into (3.170), we obtain

$$\sigma^2_{\epsilon_\vartheta} \leq 16\frac{\hat{g}^2}{\sigma^2_\mathrm{n}\gamma}\lambda\hat{H}^2_\varsigma\, \mathrm{e}^{-\nu\rho} = 16\frac{\hat{g}^2}{\nu^2}\lambda\, \mathrm{e}^{-\nu\rho}, \quad \forall\lambda \geq 1, \tag{3.189}$$

where (3.15) and (3.169) have been used. This completes the proof. □

Thus, we have obtained an upper bound on the MSE due to clipping errors that grows linearly with $\lambda$ and decays exponentially with $\rho$ (provided the product $\gamma\lambda$ does not tend to zero as $\lambda \to \infty$, see (3.169)).

It is worth noting that the above bound is not tight, which stems from the use of the following inequalities: from (3.166) (Bernstein's inequality); from inequality (3.170) (which assumes that all the power of clipping errors coincides with the peak of $\left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|$); from inequality (3.178) (which only has the effect of introducing a constant scale factor); and from (3.183), (3.182), and (3.184) (which can be expected to be very loose inequalities).

**Remark 3.2.** *If one assumes that the average power of clipping errors $\sigma^2_\vartheta$ is evenly distributed over $[-\pi, \pi]$, then* (3.170) *would change to*

$$\sigma^2_{\epsilon_\vartheta} = \sigma^2_\vartheta \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|P(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})\right|^2 d\omega. \tag{3.190}$$

*From* (3.175)*, and because $g(\omega)$ is zero for $|\omega| > \pi$, the integral in* (3.190) *converges, as $\lambda \to \infty$, to the value $1/(2\sigma^2_\mathrm{n}(\gamma+1))$, which is independent of $\lambda$. This, when substituted into* (3.189)*, would eliminate the $\lambda$ factor that multiplies the exponential in* (3.168)*, yielding*

$$\sigma^2_{\epsilon_\vartheta} \leq \frac{16}{\nu}\, \mathrm{e}^{-\nu\rho}\,. \tag{3.191}$$

▲

Now we can upper bound the total MSE:

**Theorem 3.15.** *Suppose there exists a scalar $\hat{g} < \infty$ such that $g_1(\omega) \leq \hat{g},\ \forall \omega \in [-\pi, \pi]$, see* (3.154)*. Assume that the innovations of the process $\{\mathrm{x}(k)\}$ is a sequence of zero-mean independent random variables $\{\xi(k)\}$ having a symmetric PDF and moments which satisfy* (3.165) *with $H = H_\xi$, for some constant $H_\xi$. If the loading factor in an optimal PRFQ with an $N$-level uniform quantizer using clipping and subtractive dither varies with the oversampling ratio $\lambda$ as*

$$\rho = 4^{-1/3}\sqrt{3}(N-1)^{2/3}\lambda^{1/3}, \tag{3.192}$$

*then $\sigma_\epsilon^2$, the MSE including overload errors, satisfies*

$$\sigma_\epsilon^2 = \mathcal{O}(\mathrm{e}^{-c_0\lambda^{1/3}}), \quad \text{as } \lambda \to \infty, \tag{3.193}$$

*where the constant*

$$c_0 \triangleq [0.5(N-1)]^{2/3}. \tag{3.194}$$

▲

*Proof.* The total frequency weighted error is

$$\epsilon = \epsilon_n + \epsilon_\vartheta, \tag{3.195}$$

where

$$\epsilon_n(k) \triangleq (1 - F(z))B(z)n(k) \tag{3.196}$$

is the term in $\epsilon$ due to granular errors in $\mathcal{Q}$ and where $\epsilon_\vartheta$ is the part of $\epsilon$ due to clipping errors. We then have that

$$\sigma_\epsilon^2 = \mathrm{E}\left[\epsilon^2\right] = \mathrm{E}\left[(\epsilon_n + \epsilon_\vartheta)^2\right] \leq \mathrm{E}\left[2\left(\epsilon_n^2 + \epsilon_\vartheta^2\right)\right] = 2(\sigma_{\epsilon_n}^2 + \sigma_{\epsilon_\vartheta}^2). \tag{3.197}$$

By substituting (3.177) into (3.159), the upper bound to the FWMSE due to granular quantization errors in (3.159) becomes

$$\sigma_{\epsilon_n}^2 \leq \hat{g}^2 \left(\frac{\gamma+1}{\gamma}\right)^3 \mathrm{e}^{-\ln(\gamma+1)\lambda} \tag{3.198}$$

Upon substituting (3.189) and (3.168) in (3.197), we obtain the following upper bound for the total error variance:

$$\sigma_\epsilon^2 \leq 2\hat{g}^2 \left(\frac{\gamma+1}{\gamma}\right)^3 \mathrm{e}^{-\ln(\gamma+1)\lambda} + 32\frac{\hat{g}^2}{\nu^2}\lambda\,\mathrm{e}^{-\nu\rho}. \tag{3.199}$$

The above upper bound for $\sigma_\epsilon$ does not tend to zero with increasing $\lambda$ unless one makes the loading factor $\rho$ grow with $\lambda$ fast enough. From (3.162) and since the use of subtractively dithered uniform scalar quantization reduces the effective number of quantization levels by 1, we have that $\gamma = \eta/\rho^2$, where $\eta \triangleq 3(N-1)^2$. Thus, the term due to clipping errors in (3.199) can be reduced only at the expense of having $\mathcal{Q}$ operate at a lower SNR. This, in turn, makes the term due to granular errors decay more slowly with increasing $\lambda$.

For example, if one makes the loading factor $\rho$ grow with $\lambda$ as $\rho = \varpi\lambda^p$, where $p > 0$ and $a > 0$ are constants to be chosen, then (3.199) becomes

$$\sigma_\epsilon^2 \leq 2\hat{g}^2 \left(1 + \frac{\varpi^2\lambda^{2p}}{\eta}\right)^3 \mathrm{e}^{-\ln(\frac{\eta}{\varpi^2}\lambda^{-2p}+1)\lambda} + 32\frac{\hat{g}^2}{\nu^2}\lambda\,\mathrm{e}^{-\nu\varpi\lambda^p}. \tag{3.200}$$

It can be seen in the above expression that a large value of $p$ would reduce the decay of the granular error term and increase that of the clipping error term, as $\lambda$ grows. Thus, the optimal decay rate when $\lambda \to \infty$ is achieved by choosing $p$ so as to make both terms decay at the same asymptotic rate. This is achieved if and only if $p$ and $a$ are chosen so that

$$c \triangleq \lim_{\lambda\to\infty} \frac{\ln\left(\frac{\eta}{\varpi^2\lambda^{2p}}+1\right)\lambda - 3\ln\left(1+\frac{\varpi^2\lambda^{2p}}{\eta}\right) - \ln 2}{\nu\varpi\lambda^p - \ln(\lambda) - 2\ln(4\hat{g}/\nu) - \ln 2} \tag{3.201}$$

equals 1. Before evaluating the above limit, note that from (3.169) we obtain

$$\nu = \breve{\nu} \triangleq 4\sqrt{3}, \quad \forall\lambda \geq 2\sqrt{2}\left(\frac{\gamma+1}{\gamma}\right)\frac{H_\xi^2}{\sigma_\xi^2}, \tag{3.202}$$

since $n$, being a random variable uniformly distributed over $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$, has standard deviation $\sigma_{\mathrm{n}} = \frac{\Delta}{2\sqrt{3}}$ and satisfies (3.165) with $H_n = \frac{\Delta}{8}$. Applying l'Hôpital's rule to (3.201) and substituting $\nu$ by $\breve{\nu}$,

$$\begin{aligned}
c &= \lim_{\lambda\to\infty}\left[\frac{\ln\left(\frac{\eta}{\varpi^2}\lambda^{-2p}+1\right)[\lambda-3] - 6p\ln(\lambda) + 3\ln\left(\frac{\eta}{\varpi^2}\right)}{\breve{\nu}\varpi\lambda^p - \ln(\lambda) - \ln(16/3)}\right] \\
&= \lim_{\lambda\to\infty}\left[\frac{\ln\left(\frac{\eta}{\varpi^2}\lambda^{-2p}+1\right) - \frac{2\eta p}{\eta+\varpi^2\lambda^{2p}}[1-3\lambda^{-1}] - 6p\lambda^{-1}}{\breve{\nu}\varpi p\lambda^{p-1} - \lambda^{-1}}\right] \\
&= \lim_{\lambda\to\infty}\left[\frac{\frac{-2\eta p\lambda^{-1}}{\eta+\varpi^2\lambda^{2p}} - \frac{6\eta p\lambda^{-2}(\eta+\varpi^2\lambda^{2p}) - 4\varpi^2\eta p^2[1-3\lambda^{-1}]\lambda^{2p-1}}{(\eta+\varpi^2\lambda^{2p})^2} + 6p\lambda^{-2}}{\breve{\nu}\varpi p(p-1)\lambda^{p-2} + \lambda^{-2}}\right] \\
&= \lim_{\lambda\to\infty}\left[\frac{\frac{-2\eta p\lambda}{\eta+\varpi^2\lambda^{2p}} - \frac{6\eta p(\eta+\varpi^2\lambda^{2p}) - 4\varpi^2\eta p^2[1-3\lambda^{-1}]\lambda^{2p+1}}{(\eta+\varpi^2\lambda^{2p})^2} + 6p}{\breve{\nu}\varpi p(p-1)\lambda^p + 1}\right] \\
&= \lim_{\lambda\to\infty}\left[\frac{-2\eta p\lambda(\eta+\varpi^2\lambda^{2p}) - 6\eta p(\eta+\varpi^2\lambda^{2p}) + 4\varpi^2\eta p^2[1-3\lambda^{-1}]\lambda^{2p+1} + 6p(\eta+\varpi^2\lambda^{2p})^2}{(\eta^2+\varpi^4\lambda^{4p}+2\varpi^2\eta\lambda^{2p})(\breve{\nu}\varpi p(p-1)\lambda^p+1)}\right].
\end{aligned} \tag{3.203}$$

By comparing the powers of $\lambda$ in the numerator and denominator of the right-hand side of (3.203), it is clear from that $c$ is either 0 or $\infty$ unless $p = 1/3$. With this choice, we get $c = \frac{\eta}{\breve{\nu}\varpi^3}$, and thus $c = 1 \iff \varpi = \left(\frac{\eta}{\breve{\nu}}\right)^{1/3}$. Therefore, the right-hand side of (3.201) equals 1 if and only if

$$\begin{aligned} p &= 1/3, \\ \varpi &= \left(\frac{\eta}{\breve{\nu}}\right)^{1/3}. \end{aligned}$$

Substituting these values into (3.200) we obtain

$$\sigma_\varepsilon^2 \leq \mathrm{e}^{-h_1(\lambda)} + \mathrm{e}^{-h_2(\lambda)}, \tag{3.204}$$

where

$$\begin{aligned} h_1(\lambda) &\triangleq \ln\left[c_1\lambda^{-2/3} + 1\right]\lambda - 3\ln\left[1 + c_1^{-1}\lambda^{2/3}\right] - \ln(2\hat{g}^2), \\ h_2(\lambda) &\triangleq c_1\lambda^{1/3} - \ln(\lambda) - \ln(32\hat{g}^2\nu^{-2}), \end{aligned} \tag{3.205}$$

$\forall\lambda > 1$, and where

$$c_1 \triangleq \eta^{1/3}\breve{\nu}^{2/3} = (3/16)^{1/3}\eta^{1/3} = (3/16)^{1/3}[3(N-1)^2]^{1/3} = [(3/4)(N-1)]^{2/3}. \tag{3.206}$$

From (3.205) and (3.206), it is straightforward to show that

$$\lim_{\lambda\to\infty}\frac{h_1(\lambda)}{\lambda^{1/3}} = \lim_{\lambda\to\infty}\frac{h_2(\lambda)}{\lambda^{1/3}} = c_1. \tag{3.207}$$

Using (3.207) it is found that, for any constant $c < c_1$, the following holds

$$\lim_{\lambda\to\infty}\left(\frac{\mathrm{e}^{-h_1(\lambda)}}{\mathrm{e}^{-c\lambda^{1/3}}}\right)^{\frac{1}{c\lambda^{1/3}}} = \lim_{\lambda\to\infty}\mathrm{e}^{1-\frac{h_1(\lambda)}{c\lambda^{1/3}}} = \mathrm{e}^{1-\frac{c_1}{c}} < 1. \tag{3.208}$$

This means that for every $\varepsilon > 0$, there exists a bounded and positive $\Lambda_1 = \Lambda_1(\varepsilon)$ such that, if $\lambda > \Lambda_1$, then $\left(\frac{\mathrm{e}^{-h_1(\lambda)}}{\mathrm{e}^{-c\lambda^{1/3}}}\right)^{\frac{1}{c\lambda^{1/3}}} \leq \mathrm{e}^{1-\frac{c_1}{c}} + \varepsilon$. Choosing $\varepsilon < 1 - \mathrm{e}^{1-\frac{c_1}{c}}$, we have that

$$\left(\frac{\mathrm{e}^{-h_1(\lambda)}}{\mathrm{e}^{-c\lambda^{1/3}}}\right)^{\frac{1}{c\lambda^{1/3}}} \leq \mathrm{e}^{1-\frac{c_1}{c}} + \varepsilon < 1, \quad \forall\lambda > \Lambda_1(\varepsilon) \quad \iff \quad \frac{\mathrm{e}^{-h_1(\lambda)}}{\mathrm{e}^{-c\lambda^{1/3}}} < 1, \quad \forall\lambda > \Lambda_1(\varepsilon). \tag{3.209}$$

A similar analysis leads to the existence of a bounded and positive $\Lambda_2(\varepsilon)$ such that

$$\frac{\mathrm{e}^{-h_2(\lambda)}}{\mathrm{e}^{-c\lambda^{1/3}}} < 1, \quad \forall\lambda > \Lambda_2(\varepsilon). \tag{3.210}$$

Since $c_0$ in (3.194) satisfies $c_0 < c_1$, (3.209) and (3.210) demonstrate (3.193). This completes the proof. □

**Remark 3.3.** *The requirement that the innovation process of $\{\mathrm{x}(k)\}$ is made of independent samples is necessary in Theorem 3.15 in order to obtain* (3.184) *in Theorem 3.14. This condition can be omitted if the pre-filter is simply a scalar gain and if we assume, instead, that the samples of $\{\mathrm{x}(k)\}$ satisfy* (3.165) *for some bounded $H$. Setting the pre-filter as a scalar gain would yield a non optimal PRFQ (unless, of course, the source is white). Nevertheless, one would obtain that $\sigma^2_\vartheta$ also grows linearly with $\lambda$ and decays exponentially with $\rho$. As a consequence, Theorem 3.15 can be extended in a similar fashion without requiring the source to have an innovations process with independent samples.*

**Remark 3.4.** *From Remark 3.2, if one makes the assumption that the average spectral power of $\vartheta$ is evenly distributed over $[-\pi, \pi]$, the effect would be to eliminate the $+1$ term on the right end of the denominator of* (3.203)*. Following the proof of Theorem 3.15, it is easy to verify that this* would have no effect *on* (3.193)*.*

## 3.13 Summary

In this chapter we have characterized the filters around a scalar quantizer with given SNR that minimize the frequency weighted reconstruction WCMSE. The associated optimal performance (SNR-distortion) trade-off for this class of ED pairs has been also established. It has been shown that the frequency weighted MSE of optimal perfect reconstruction feedback quantizers decreases exponentially with the oversampling ratio, if the quantizer SNR is kept constant. In addition, a lower bound to this decay ratio has been found when the number of levels in the quantizer is finite and fixed. This bound takes into account the effect of clipping errors, and holds for sources with unbounded support.

## 3.14 Appendix

In this appendix we give some technical results which were used throughout this chapter.

**Lemma 3.16** (WCMSE Wiener Filter)**.** *Let $S_{\mathrm{n}}(\mathrm{e}^{j\omega}) = \left|\Omega_{\mathrm{n}}(\mathrm{e}^{j\omega})\right|^2$ and $S_{\mathrm{x}} = \left|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})\right|^2$ be power spectral densities. Let $a, b > 0$ be given constants. Then, the frequency response of the LTI filter $W(z)$ that minimizes*

$$D_{a,b}(\mathrm{x}, \mathrm{y}) = a\|W\Omega_{\mathrm{n}}P\|^2 + b\|(W-1)\Omega_{\mathrm{x}}P\|^2 \tag{3.211}$$

*satisfies*

$$W(\mathrm{e}^{j\omega}) = \frac{bS_{\mathrm{x}}(\mathrm{e}^{j\omega})}{aS_{\mathrm{n}}(\mathrm{e}^{j\omega}) + bS_{\mathrm{x}}(\mathrm{e}^{j\omega})}, \quad \text{a.e. on } [-\pi, \pi] \setminus \mathcal{N}_P, \tag{3.212}$$

*where* $\mathcal{N}_P \triangleq \{w \in [-\pi, \pi] : P(\mathrm{e}^{j\omega}) = 0\}$. ▲

*Proof.* It is clear from (3.211) that in order to minimize $D_{a,b}$, $W(\mathrm{e}^{j\omega})$ must be real, non-negative and symmetric. If this is the case, we have that

$$
\begin{aligned}
&D_{a,b}(\mathrm{x},\mathrm{y}) \\
&= \int \left|P(\mathrm{e}^{j\omega})\right|^2 \left(aW(\mathrm{e}^{j\omega})^2\Omega_{\mathrm{n}}(\mathrm{e}^{j\omega})^2 + bW(\mathrm{e}^{j\omega})^2\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2 - 2bW(\mathrm{e}^{j\omega})\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2 + b\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2\right) d\omega \\
&= \int \left|P(\mathrm{e}^{j\omega})\right|^2 \left(\left(a\Omega_{\mathrm{n}}^2 + b\Omega_{\mathrm{x}}^2\right) W^2 - 2bW\Omega_{\mathrm{x}}^2 + b\Omega_{\mathrm{x}}^2\right) d\omega \\
&= \int \left|P(\mathrm{e}^{j\omega})\right|^2 \left(\left[\sqrt{a\Omega_{\mathrm{n}}^2 + b\Omega_{\mathrm{x}}^2}W - \frac{b\Omega_{\mathrm{x}}^2}{\sqrt{a\Omega_{\mathrm{n}}^2 + b\Omega_{\mathrm{x}}^2}}\right]^2 + b\Omega_{\mathrm{x}}^2 d\omega - \left[\frac{b\Omega_{\mathrm{x}}^2}{\sqrt{a\Omega_{\mathrm{n}}^2 + b\Omega_{\mathrm{x}}^2}}\right]^2\right) d\omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|P(\mathrm{e}^{j\omega})\right|^2 T(\mathrm{e}^{j\omega}) \left[W(\mathrm{e}^{j\omega}) - \frac{b\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2}{T(\mathrm{e}^{j\omega})^2}\right] + \left|P(\mathrm{e}^{j\omega})\right|^2 \frac{ab\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2\Omega_{\mathrm{n}}(\mathrm{e}^{j\omega})^2}{T(\mathrm{e}^{j\omega})^2} d\omega \\
&= \left\| P\left(W - \frac{b\Omega_{\mathrm{x}}^2}{T^2}\right) T \right\|^2 + \left\| P\frac{\sqrt{ab}\Omega_{\mathrm{x}}\Omega_{\mathrm{n}}}{T} \right\|^2
\end{aligned}
\tag{3.213}
$$

where $T(\mathrm{e}^{j\omega}) \triangleq \sqrt{a\Omega_{\mathrm{n}}(\mathrm{e}^{j\omega})^2 + b\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2}$. From the last line of (3.213), we conclude that the filter $W(z)$ that minimizes $D_{a,b}(\mathrm{x},\mathrm{y})$ has the frequency response given in (3.212). □

As one could expect from (3.211), the filter described by (3.212) *would* be the standard non-causal Wiener filter if the source and the noise had power spectral densities $a\left|P(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega})$ and $b\left|P(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{n}}(\mathrm{e}^{j\omega})$, respectively.

**Theorem 3.17** (Simplified from Theorem 1 on p. 217 of [128])**.** *Let $\mathcal{X}$ be a linear vector space and $\mathcal{S}$ a convex subset of $\mathcal{X}$. Let $\mathscr{F}$ be a real-valued convex functional on $\mathcal{S}$ and $\mathscr{G}_1, \mathscr{G}_2, \ldots, \mathscr{G}_N$ convex mappings from $\mathcal{S}$ into $\mathbb{R}$. Assume the existence of a point $f_1 \in \mathcal{S}$ such that $\mathscr{G}_i(f_1) < 0$, for $i = 1, 2, \ldots, N$.*

*Let*

$$
\mu_0 \triangleq \inf \mathscr{F}(f) \quad \textit{subject to } f \in \mathcal{S}, \mathscr{G}_i(f) \leq 0, i = 1, 2, \ldots, N \tag{3.214}
$$

*and assume $\mu_0$ is finite. Then there exists a vector $\boldsymbol{\lambda} \in \mathbb{R}^N$ satisfying*

$$
\lambda_i \geq 0, \quad \forall i \in \{1, 2, \ldots, N\} \tag{3.215}
$$

*and such that*

$$
\mu_0 = \inf_{f \in \mathcal{S}} \left\{ \mathscr{F}(f) + \sum\nolimits_{i=1}^{N} \lambda_i \mathscr{G}_i(f) \right\}, \tag{3.216}
$$

*Furthermore, if the infimum is achieved in* (3.214) *by* $f^\star \in \mathcal{S}$, $\mathscr{G}_i(f^\star) \leq 0, \forall i \in \{1, 2, \ldots, N\}$, *then* $f^\star$ *also achieves the infimum in* (3.216) *and*

$$\mathscr{G}_i(f^\star)\lambda_i = 0, \quad , \forall i \in \{1, 2, \ldots, N\} \tag{3.217}$$

*Proof.* Same as in the proof of [128, Theorem 1, in p. 217]. □

**Theorem 3.18.** [15] *If* $\phi, \psi : [a, b] \to \mathbb{R}$ *are similarly functionally related, then*

$$[b-a]\int_a^b \phi(x)\psi(x)dx \geq \int_a^b \phi(x)dx \int_a^b \psi(x)dx. \tag{3.218}$$

*If* $\phi$ *and* $\psi$ *are oppositely functionally related, then the inequality in* (3.218) *is reversed. In either case, equality is achieved if and only if* $\psi$ *(and therefore* $\phi$*) is almost constant.* ▲

*Proof.* We will examine the difference between the right and left hand side in (3.218). We obtain

$$\int_a^b \phi(x)\psi(x)dx - \overline{\psi}\int_a^b \phi(x)dx = \int_a^b \phi(x)\left[\psi(x) - \overline{\psi}\right]dx,$$

where $\overline{\psi} \triangleq \frac{1}{b-a}\int_a^b \psi(x)dx$. Note that we have divided both sides by $b - a$. Suppose $\phi \uparrow\uparrow \psi$. (The proof for $\phi \uparrow\downarrow \psi$ proceeds in a similar way.) Then there exists a monotonically increasing function $G(\cdot)$ such that $\phi = G(\psi)$, and a value $\phi_0$ such that $\phi(x) > \phi_0 \iff \psi(x) > \overline{\psi}$ and $\phi(x) < \phi_0 \iff \psi(x) < \overline{\psi}$. It then follows that

$$\int_a^b \phi(x)\left[\psi(x) - \overline{\psi}\right]dx \geq \int_a^b \phi_0\left[\psi(x) - \overline{\psi}\right]dx = 0,$$

with equality if and only if

$$\int_{\psi > \overline{\psi}} [\psi(x) - \overline{\psi}]dx = 0 = \int_{\psi < \overline{\psi}} [\psi(x) - \overline{\psi}]dx,$$

i.e., if and only if $\psi$ (and therefore $\phi$ as well) is almost constant. □

[15] This theorem is related to the variant of Tchebyshev's Integral Inequality given in [150, Theorem 236]. It departs from the latter in that the integrands must be functionally dependent, which allows us to state necessary and sufficient conditions for equality.

# Chapter 4

# The WCMSE-Rate-Distortion Function

*It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.*

*Albert Einstein*

*There is nothing more practical than a good theory.*
*Attributed to Kurt Lewin, German Psychologist.*

## 4.1 Introduction

In the previous chapter we analyzed the optimal bit-rate performance attainable with a general feedback scalar quantization scheme, using WCMSE as the distortion metric. However, it is not clear from those results whether the achieved performance is optimal in an absolute sense, that is, when compared to the best performance achievable by *any* possible source coding scheme. A function that characterizes the minimum achievable bit-rate for a given distortion metric is, by definition, a rate-distortion function [6].

Here, we first define the information-theoretic rate-distortion function for WCMSE as the distortion metric, denoted as WCMSE-RDF or by and by the function $R_{a,b}(D)$. The information-theoretic WCMSE-RDF is then characterized for Gaussian sources, which serves, initially, as a lower bound to the rate attainable by any ED pair under the constraint that the WCMSE is smaller than some value $D > 0$. The cases of Gaussian scalar and vector sources are discussed in sections 4.3, 4.4 respectively. The information-theoretic WCMSE-RDF for Gaussian stationary scalar processes and vector processes

is characterized in sections 4.5 and 4.6, respectively. A proof of achievability of these bounds is provided in Section 4.8, which implies that the information-theoretic WCMSE rate-distortion function actually coincides with the WCMSE rate-distortion function. The achievability result holds for Gaussian scalar and vector processes, as well as for infinite ensembles of Gaussian scalar random variables.

The special cases in which the WCMSE has weights $a = b = 1$ and $a = 1$, $b = \infty$ are briefly discussed in Section 4.5.2. The latter case characterizes the *quadratic Gaussian rate-distortion function for source-uncorrelated distortions*, denoted by $R^{\perp}(D)$ and recently introduced by the author and colleagues in [127]. An image processing example which provides a tangible illustration of the meaning and potential applicability of the WCMSE-RDF is presented in Section 4.7. In Section 4.9, this rate-distortion function is extended and characterized for cases in which there exists linear, time-invariant (LTI) feedback between reconstruction and source.

We begin by stating some preliminary results.

## 4.2 Preliminaries

### 4.2.1 The WCMSE is Not Linear in the PDF of Source and Reconstruction

Perhaps surprisingly, the WCMSE (defined in Section 1.3.1) differs from the standard MSE in an essential way. It is clear that both WCMSE and MSE are expectation (or ensemble average) distortion metrics. However, unlike MSE, the WCMSE *cannot be generated from a fidelity criterion*, in the strict sense of the term. This stands in stark contrast with most distortion metrics studied in the rate-distortion theory literature, see, e.g., [6].

To demonstrate this fact, we first recall that, for a single, real valued scalar source $\mathrm{x}$, a (scalar) *distortion measure* is a function

$$\rho(x, y), \quad \rho : \mathbb{R}^2 \to \mathbb{R}_0^+,$$

that represents the cost of having a realization $x$ of the source being reconstructed as $y$. In the scalar case, a distortion metric $D(\mathrm{x}, \mathrm{y})$ is usually generated from a given measure by taking the expected value of $\rho(\mathrm{x}, \mathrm{y})$, that is, $D(\mathrm{x}, \mathrm{y}) = \mathrm{E}\left[\rho(\mathrm{x}, \mathrm{y})\right]$. In such cases, $D(\mathrm{x}, \mathrm{y})$ is simply called a distortion.

Crucially, when originated as the expected value of a distortion measure, a distortion metric is linear in the joint probability distribution of source, $\mathrm{x}$, and reconstruction, $\mathrm{y}$. For example, consider a given (scalar) distortion measure $\rho(x, y)$ and a scalar source $\mathrm{x}$ with PDF $f_{\mathrm{x}}(\cdot)$. For this source, two conditional probability assignments $f_{\mathrm{y}_1 \,|\, \mathrm{x}}(\cdot|\cdot)$, $f_{\mathrm{y}_2 \,|\, \mathrm{x}}(\cdot|\cdot)$ generate two values for the distortion metric $D(\mathrm{x}, \mathrm{y}) =$

$\mathrm{E}\left[\rho(\mathrm{x},\mathrm{y})\right]$, given by

$$D_1 = \iint \rho(x,y) f_{\mathrm{y}_1\,|\,\mathrm{x}}(y|x) f(x) dy dx$$
$$D_2 = \iint \rho(x,y) f_{\mathrm{y}_2\,|\,\mathrm{x}}(y|x) f(x) dy dx,$$

respectively. Then, for the conditional PDF assignment $f_{\mathrm{y}_3\,|\,\mathrm{x}}(y|x) = c f_{\mathrm{y}_1\,|\,\mathrm{x}}(y|x) + (1-c) f_{\mathrm{y}_2\,|\,\mathrm{x}}(y|x)$, with $0 \leq c \leq 1$, the average distortion is

$$\begin{aligned} D_3 &= \iint \rho(x,y) \left[ c f_{\mathrm{y}_1\,|\,\mathrm{x}}(y|x) + (1-c) f_{\mathrm{y}_2\,|\,\mathrm{x}}(y|x) \right] f(x) dy dx \\ &= c \iint \rho(x,y) f_{\mathrm{y}_1\,|\,\mathrm{x}}(y|x) f(x) dy dx + (1-c) \iint \rho(x,y) f_{\mathrm{y}_2\,|\,\mathrm{x}}(y|x) f(x) dy dx \\ &= c D_1 + (1-c) D_2. \end{aligned}$$

This example illustrates the fact that, if a distortion metric is expected value of a distortion measure, then the distortion obtained by a weighted mixture of two or more operating regimes is the weighted sum of the distortions associated with each regime.

To demonstrate that the WCMSE cannot be expressed as the expected value of a distortion measure, it suffices to give an example in which the linearity of the WCMSE with respect to the probability distribution of source and reconstruction does not hold. For this purpose, consider a hybrid operating regime, obtained as the combination of two basic operating regimes. When the regime indicator variable $r$ equals 1, which occurs with probability $c$, the reconstruction error is given by $\mathrm{z}_1 = \mathrm{u} - V\,\mathrm{x}$, where $V$ is a scalar and $\mathrm{u}$ is a random variable uncorrelated to $\mathrm{x}$. On the other hand, when $r$ equals 2, which occurs with probability $(1-c)$, the reconstruction error is $\mathrm{z}_2 = -\mathrm{x}$. The reconstruction error obtained from the stochastic combination of these two regimes is characterized via

$$\mathrm{z}_3 = \begin{cases} \mathrm{z}_1 & , \text{ if } r = 1, \\ \mathrm{z}_2 & , \text{ if } r = 2. \end{cases}$$

The source-parallel error in the hybrid regime is $(\sigma_{\mathrm{x},\mathrm{z}_3}/\sigma_{\mathrm{x}}^2)\,\mathrm{x}$. The covariance between $\mathrm{x}$ and $\mathrm{z}_3$ is given by

$$\begin{aligned} \sigma_{\mathrm{x},\mathrm{z}_3} &= \mathrm{E}\left[\mathrm{x}\,\mathrm{z}_3\right] = \mathrm{E}\left[\mathrm{x}\,\mathrm{z}_3\,|r=1\right] c + \mathrm{E}\left[\mathrm{x}\,\mathrm{z}_3\,|r=2\right](1-c) \\ &= \mathrm{E}\left[\mathrm{x}(\mathrm{u} - V\,\mathrm{x})\right] c + \mathrm{E}\left[\mathrm{x}(-\mathrm{x})\right](1-c) = -V\sigma_{\mathrm{x}}^2 c - \sigma_{\mathrm{x}}^2(1-c) \\ &= -\sigma_{\mathrm{x}}^2\left(Vc + 1 - c\right) \end{aligned}$$

Thus,

$$D_3^{\parallel} = \frac{\sigma_{\mathrm{x},\mathrm{z}_3}^2}{\sigma_{\mathrm{x}}^2} = \left(Vc + 1 - c\right)^2 \sigma_{\mathrm{x}}^2.$$

The source-uncorrelated distortion term for the hybrid regime is $D_3^{\perp} = \sigma^2_{\mathrm{z}_3} - D_3^{\parallel}$. We will first determine the variance of $\mathrm{z}_3$, namely

$$\begin{aligned}\sigma^2_{\mathrm{z}_3} &= \mathrm{E}\left[\mathrm{z}_3^2\right] = \mathrm{E}\left[\mathrm{z}_3^2\,|r=1\right]c + \mathrm{E}\left[\mathrm{z}_3^2\,|r=2\right](1-c)\\ &= \mathrm{E}\left[(\mathrm{u}-V\,\mathrm{x})^2\right]c + \mathrm{E}\left[\mathrm{x}^2\right](1-c)\\ &= \sigma^2_{\mathrm{u}}c + V^2\sigma^2_{\mathrm{x}}c + \sigma^2_{\mathrm{x}}(1-c). \end{aligned} \tag{4.1}$$

With this, and recalling that $D^{\perp} = \sigma^2_{\mathrm{z}} - D^{\parallel}$, we obtain

$$D_3^{\perp} = \sigma^2_{\mathrm{u}}c + \sigma^2_{\mathrm{x}}\left[\left(V^2c+1-c\right) - \left(Vc+1-c\right)^2\right].$$

Notice from 4.1 that $\sigma^2_{\mathrm{z}_3} = c\sigma^2_{\mathrm{z}_1} + (1-c)\sigma^2_{\mathrm{z}_2}$, i.e., the hybrid MSE is the linear combination of the MSEs of each regime. However, for the WCMSE,

$$\begin{aligned}D_{a,b}(\mathrm{x},\mathrm{x}+\mathrm{z}_3) &= a\sigma^2_{\mathrm{u}}c + a\sigma^2_{\mathrm{x}}\left[\left(V^2c+1-c\right) - \left(Vc+1-c\right)^2\right] + b\left(Vc+1-c\right)^2\sigma^2_{\mathrm{x}}\\ &= a\left\{\sigma^2_{\mathrm{u}}c + \sigma^2_{\mathrm{x}}\left[\left(V^2c+1-c\right) - \left(Vc+1-c\right)^2\right]\right\} + b\left(Vc+1-c\right)^2\sigma^2_{\mathrm{x}}\\ &\neq a\sigma^2_{\mathrm{u}}c + bV^2\sigma^2_{\mathrm{x}}c + b\sigma^2_{\mathrm{x}}(1-c) = cD_{a,b}(\mathrm{x},\mathrm{x}+\mathrm{z}_1) + (1-c)D_{a,b}(\mathrm{x},\mathrm{x}+\mathrm{z}_2),\end{aligned}$$

where equality holds if and only if $a = b$ or $V = 1$, i.e., in the special case in which the WCMSE is a scaled version of the standard MSE, or when the first regime equals the second.

The above example proves that WCMSE, in general, cannot be expressed as the expected value of a distortion measure. In particular, $D_{a,b}$ cannot be expressed as the expected value of a difference distortion measure of the form $\rho(\boldsymbol{y}-\boldsymbol{x})$, which is the focus in, e.g., [6, 151], and $D_{a,b}$ cannot be expressed as the expected value of an input-dependent distortion measure of the form $\rho_{\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{x})$ or $\rho(\boldsymbol{x},\boldsymbol{y})$, such as those studied in, e.g., [18, 152, 153]. Nevertheless, it is still possible to characterize the rate-distortion function for Gaussian sources using the WCMSE as the distortion metric, as will be shown in Section 4.3 below.

### 4.2.2 The Reconstruction Error Must Be Jointly Gaussian with the Source

The next lemma plays an important role in the results derived subsequently in this section. It will be used to conclude that, for every realization of $R_{a,b}(D)$, the reconstruction error is necessarily jointly Gaussian with the source.

**Lemma 4.1.** *Let $\mathbf{x} \in \mathbb{R}^N \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}_{\mathbf{x}})$, with $|\boldsymbol{K}_{\mathbf{x}}| > 0$. Let $\mathbf{z} \in \mathbb{R}^N$ and $\mathbf{z}_G \in \mathbb{R}^N$ be two random vectors with zero mean and the same covariance matrix, i.e., $\boldsymbol{K}_{\mathbf{z}} = \boldsymbol{K}_{\mathbf{z}_G}$, also having the same cross-covariance matrix with respect to $\mathbf{x}$, that is, $\boldsymbol{K}_{\mathbf{x},\mathbf{z}} = \boldsymbol{K}_{\mathbf{x},\mathbf{z}_G}$. If $\mathbf{z}_G$ and $\mathbf{x}$ are jointly Gaussian, and if $\mathbf{z}$*

*has any distribution, then the mutual information between* $\mathbf{x}$ *and* $\mathbf{x}+\mathbf{z}$ *satisfies*

$$I(\mathbf{x};\mathbf{x}+\mathbf{z}) \geq I(\mathbf{x};\mathbf{x}+\mathbf{z}_G). \tag{4.2}$$

*Equality is achieved in* (4.2) *if and only if* $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}_{\mathbf{z}})$ *with* $\mathbf{z}$ *and* $\mathbf{x}$ *being jointly Gaussian.*

*Proof.* Define $\mathbf{y} \triangleq \mathbf{x}+\mathbf{z}$ and $\mathbf{y}_G \triangleq \mathbf{x}+\mathbf{z}_G$. Then

$$\begin{aligned}
&I(\mathbf{x};\mathbf{x}+\mathbf{z}) - I(\mathbf{x};\mathbf{x}+\mathbf{z}_G) = h(\mathbf{x}|\mathbf{y}_G) - h(\mathbf{x}|\mathbf{y}) = h(\mathbf{z}_G|\mathbf{y}_G) - h(\mathbf{z}|\mathbf{y}) \\
&= -\iint f_{\mathbf{z}_G,\mathbf{y}_G}(\boldsymbol{z},\boldsymbol{y}) \log(f_{\mathbf{z}_G|\mathbf{y}_G}(\boldsymbol{z}|\boldsymbol{y}))d\boldsymbol{z}d\boldsymbol{y} + \iint f_{\mathbf{z},\mathbf{y}}(\boldsymbol{z},\boldsymbol{y}) \log(f_{\mathbf{z}|\mathbf{y}}(\boldsymbol{z}|\boldsymbol{y}))d\boldsymbol{z}d\boldsymbol{y} \qquad (4.3)\\
&\overset{(a)}{=} -\iint f_{\mathbf{z},\mathbf{y}}(\boldsymbol{z},\boldsymbol{y}) \log(f_{\mathbf{z}_G|\mathbf{y}_G}(\boldsymbol{z}|\boldsymbol{y}))d\boldsymbol{z}d\boldsymbol{y} + \iint f_{\mathbf{z},\mathbf{y}}(\boldsymbol{z},\boldsymbol{y}) \log(f_{\mathbf{z}|\mathbf{y}}(\boldsymbol{z}|\boldsymbol{y}))d\boldsymbol{z}d\boldsymbol{y} \\
&= \int f_{\mathbf{y}}(\boldsymbol{y}) \int f_{\mathbf{z}|\mathbf{y}}(\boldsymbol{z}|\boldsymbol{y}) \log\left(\frac{f_{\mathbf{z}|\mathbf{y}}(\boldsymbol{z}|\boldsymbol{y})}{f_{\mathbf{z}_G|\mathbf{y}_G}(\boldsymbol{z}|\boldsymbol{y})}\right) d\boldsymbol{z}d\boldsymbol{y} \\
&= \int f_{\mathbf{y}}(\boldsymbol{y}) D(f_{\mathbf{z}|\mathbf{y}=\boldsymbol{y}} \| f_{\mathbf{z}_G|\mathbf{y}_G=\boldsymbol{y}})d\boldsymbol{y} \geq 0, \qquad (4.4)
\end{aligned}$$

where $D(f\|g)$ is the relative entropy (or *Kullback-Leibler distance*) between the two probability density functions $f$ and $g$ (see Definition 2.14 on page 38). Equality $(a)$ follows from the fact that $\log(f_{\mathbf{z}_G|\mathbf{y}_G}(\boldsymbol{z}|\boldsymbol{y}))$ is a quadratic form of $\boldsymbol{z}$ and $\boldsymbol{y}$, and from the fact that $\boldsymbol{K}_{\mathbf{z},\mathbf{y}} = \boldsymbol{K}_{\mathbf{z}_G,\mathbf{y}_G}$. The inequality in (4.4) follows from the fact that $D(f\|g) \geq 0$, with equality if and only if $f = g$. Thus, equality is achieved if and only if

$$f_{\mathbf{z}_G|\mathbf{y}_G}(\boldsymbol{z}|\boldsymbol{y}) = f_{\mathbf{z}|\mathbf{y}}(\boldsymbol{z}|\boldsymbol{y}), \quad \forall \boldsymbol{z}, \forall \boldsymbol{y} \text{ such that } f_{\mathbf{y}}(\boldsymbol{y}) > 0. \tag{4.5}$$

It will be shown next that (4.5) implies that $\mathbf{z}$ and $\mathbf{x}$ are jointly Gaussian. For this purpose, first notice that (4.5), together with the fact that $\mathbf{x} = \mathbf{y} - \mathbf{z}$, implies

$$f_{\mathbf{x}}(\boldsymbol{x}) = \int f_{\mathbf{z}|\mathbf{y}}(\boldsymbol{y}-\boldsymbol{x}|\boldsymbol{y}) f_{\mathbf{y}}(\boldsymbol{y})d\boldsymbol{y} = \int f_{\mathbf{z}_G|\mathbf{y}_G}(\boldsymbol{y}-\boldsymbol{x}|\boldsymbol{y}) f_{\mathbf{y}}(\boldsymbol{y})d\boldsymbol{y}, \quad \forall \boldsymbol{x} \in \mathbb{R}^N. \tag{4.6}$$

But since $\mathbf{x}$ can also be written as $\mathbf{x} = \mathbf{y}_G - \mathbf{z}_G$, the following holds as well:

$$f_{\mathbf{x}}(\boldsymbol{x}) = \int f_{\mathbf{z}_G|\mathbf{y}_G}(\boldsymbol{y}-\boldsymbol{x}|\boldsymbol{y}) f_{\mathbf{y}_G}(\boldsymbol{y})d\boldsymbol{y}, \quad \forall \boldsymbol{x} \in \mathbb{R}^N. \tag{4.7}$$

Equating (4.7) and (4.6) yields

$$\int f_{\mathbf{z}_G|\mathbf{y}_G}(\boldsymbol{y}-\boldsymbol{x}|\boldsymbol{y}) f_{\mathbf{y}_G}(\boldsymbol{y})d\boldsymbol{y} = \int f_{\mathbf{z}_G|\mathbf{y}_G}(\boldsymbol{y}-\boldsymbol{x}|\boldsymbol{y}) f_{\mathbf{y}}(\boldsymbol{y})d\boldsymbol{y}, \quad \forall \boldsymbol{x} \in \mathbb{R}^N. \tag{4.8}$$

From the fact that $\mathbf{z}_G$ and $\mathbf{y}_G$ are jointly normally distributed, it follows that $f_{\mathbf{z}_G|\mathbf{y}_G}(\boldsymbol{z}|\boldsymbol{y})$ is the PDF of a normally distributed random vector, say $\mathbf{u}$, with fixed variance and mean $\boldsymbol{K}_{\mathbf{x}_G,\mathbf{y}_G}\boldsymbol{K}^{\dagger}_{\mathbf{y}_G}\boldsymbol{y} =$

$\boldsymbol{K}_{\mathbf{x},\mathbf{y}}\boldsymbol{K}_{\mathbf{y}}^{\dagger}\boldsymbol{y}$, $\forall \boldsymbol{y} \in \mathcal{N}_{\boldsymbol{K}_{\mathbf{y}}}^{\perp}$, see, e.g., [154, § 22.1]. This, together with the fact that $f_{\mathbf{y}}(\boldsymbol{y}) = f_{\mathbf{y}_G}(\boldsymbol{y}) = 0, \forall \boldsymbol{z}, \forall \boldsymbol{y} \notin \mathcal{N}_{\boldsymbol{K}_{\mathbf{y}}}^{\perp}$ [1], allows one to write (4.8) as

$$\int f_{\mathbf{u}}([\boldsymbol{I} - \boldsymbol{K}_{\mathbf{x},\mathbf{y}}\boldsymbol{K}_{\mathbf{y}}^{\dagger}]\boldsymbol{y} - \boldsymbol{x}) f_{\mathbf{y}_G}(\boldsymbol{y}) d\boldsymbol{y}, = \int f_{\mathbf{u}}([\boldsymbol{I} - \boldsymbol{K}_{\mathbf{x},\mathbf{y}}\boldsymbol{K}_{\mathbf{y}}^{\dagger}]\boldsymbol{y} - \boldsymbol{x}) f_{\mathbf{y}}(\boldsymbol{y}) d\boldsymbol{y}, \quad \forall \boldsymbol{x} \in \mathbb{R}^N.$$

The integrals in this equation are convolution integrals. From this fact, and noting that the Fourier transform of $f_{\mathbf{u}}$ is nonzero everywhere, we have that (4.5) implies $f_{\mathbf{y}}(\boldsymbol{y}) = f_{\mathbf{y}_G}(\boldsymbol{y})$, $\forall \boldsymbol{y} \in \mathbb{R}^N$. The latter and (4.5) imply $f_{\mathbf{z},\mathbf{y}}(\boldsymbol{z}, \boldsymbol{y}) = f_{\mathbf{z}_G,\mathbf{y}_G}(\boldsymbol{z}, \boldsymbol{y})$. Thus, we have shown that equality in (4.4) implies $\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{K}_z)$ with $\mathbf{z}$ and $\mathbf{x}$ being jointly Gaussian. The converse, that is, the fact that $f_{\mathbf{z},\mathbf{y}}(\boldsymbol{z}, \boldsymbol{y}) = f_{\mathbf{z}_G,\mathbf{y}_G}(\boldsymbol{z}, \boldsymbol{y})$, $\forall \boldsymbol{z}, \boldsymbol{y} \in \mathbb{R}^N$ implies equality in (4.4), can be readily verified from (4.4). This completes the proof. □

**Remark 4.1.** *We note that Lemma 4.1 generalizes Lemma II.2 in [155], by relaxing the requirement, used in [155], of having* $\mathbf{z}$ *and* $\mathbf{z}_G$ *independent of* $\mathbf{x}$*, to the requirement* $\boldsymbol{K}_{\mathbf{x},\mathbf{z}} = \boldsymbol{K}_{\mathbf{x},\mathbf{z}_G}$. ▲

We can now address the problem of characterizing $R_{a,b}(D)$ for Gaussian sources. We begin with the scalar case.

## 4.3 WCMSE-RDF for Gaussian Scalar Sources

For a Gaussian, zero mean random scalar source $\mathrm{x}$ reconstructed as $\mathrm{y}$, the reconstruction error is the random variable

$$\mathrm{z} \triangleq \mathrm{y} - \mathrm{x}\,. \tag{4.9}$$

Irrespective of how $\mathrm{y}$ is generated, $\mathrm{z}$ can *always* be decomposed into a source-uncorrelated term,

$$\mathrm{u} \triangleq \mathrm{z} - \frac{\sigma_{\mathrm{x},\mathrm{z}}}{\sigma_{\mathrm{x}}^2}\,\mathrm{x}, \tag{4.10}$$

and the remainder $\mathrm{z} - \mathrm{u} = (\sigma_{\mathrm{x},\mathrm{z}}/\sigma_{\mathrm{x}}^2)\,\mathrm{x}$, which depends linearly on $\mathrm{x}$. Notice that the random variable $\mathrm{u}$ is orthogonal to $\mathrm{x}$, i.e.,

$$\mathrm{E}\,[\mathrm{u}\,\mathrm{x}] = 0. \tag{4.11}$$

Upon defining

$$V \triangleq -\frac{\sigma_{\mathrm{x},\mathrm{z}}}{\sigma_{\mathrm{x}}^2},$$

[1] The variance of the random variable $\mathrm{u} \triangleq \boldsymbol{v}^T\mathbf{y}$, where $\boldsymbol{v} \in \mathcal{N}_{\boldsymbol{K}_{\mathbf{y}}}$, is $\mathrm{E}\left[(\boldsymbol{v}^T\mathbf{y})(\boldsymbol{v}^T\mathbf{y})^T\right] = \boldsymbol{v}^T\boldsymbol{K}_{\mathbf{y}}\boldsymbol{v} = 0$. By noting that $\mathrm{E}\left[(\boldsymbol{v}^T\mathbf{y})(\boldsymbol{v}^T\mathbf{y})^T\right] = \int (\boldsymbol{v}^T\boldsymbol{y})^2 f_{\mathbf{y}}(\boldsymbol{y}) d\boldsymbol{y}$, it follows that $f_{\mathbf{y}}(\boldsymbol{y}) = 0, \forall \boldsymbol{y} \notin \mathcal{N}_{\boldsymbol{K}_{\mathbf{y}}}^{\perp}$.

and then substituting (4.10) into (4.9), the reconstruction can be written as

$$\mathrm{y} = (1 - V)\,\mathrm{x} + \mathrm{u}\,. \tag{4.12}$$

From (4.11) and (4.12), it is clear that the source-uncorrelated and the source-parallel components of the WCMSE are given respectively by

$$D^{\perp} = \sigma_{\mathrm{u}}^2 \tag{4.13a}$$

$$D^{\parallel} = V^2 \sigma_{\mathrm{x}}^2. \tag{4.13b}$$

Thus,

$$D_{a,b}(\mathrm{x},\mathrm{y}) = aD^{\perp} + bD^{\parallel} = a\sigma_{\mathrm{u}}^2 + bV^2\sigma_{\mathrm{x}}^2. \tag{4.14}$$

This allows us to define the WCMSE RDF for scalar random sources, as follows:

**Definition 4.1** (Information-Theoretic WCMSE-RDF for Scalar Sources)**.** *The Information-Theoretic WCMSE rate-distortion function for a scalar random source is defined as*

$$R_{a,b}(D) \triangleq \min_{\mathrm{z}:\, D_{a,b}(\mathrm{x},\mathrm{x}+\mathrm{z}) \le D} I(\mathrm{x};\mathrm{x}+\mathrm{z}),$$

*where $D_{a,b}(\cdot,\cdot)$ is as defined in* (4.14)*.* ▲

The following theorem characterizes $R_{a,b}(D)$ for scalar Gaussian sources.

**Theorem 4.2** ($R_{a,b}(D)$ for Gaussian Scalar Sources)**.** *The rate-distortion function for a scalar source $\mathrm{x} \sim \mathcal{N}(0, \sigma_{\mathrm{x}}^2)$ with respect to the WCMSE distortion metric with weights $a$, $b$ is*

$$R_{a,b}(D) = \frac{1}{2}\ln\left(\max\left\{1\,,\ \frac{a\sigma_{\mathrm{x}}^2}{D} + 1 - \frac{a}{b}\right\}\right), \quad D > 0, \tag{4.15}$$

*where $D = D_{a,b}(\mathrm{x},\mathrm{y})$. A reconstruction error random variable $\mathrm{z}$ achieves $R_{a,b}(D)$ if and only if it is jointly Gaussian with $\mathrm{x}$ and*

$$D^{\perp} = \sigma_z^2 - \frac{\sigma_{\mathrm{x,z}}^2}{\sigma_{\mathrm{x}}^2}(= \sigma_{\mathrm{u}}^2) = \max\left\{0\,,\ \frac{D}{a}\left(1 - \frac{D}{b\sigma_{\mathrm{x}}^2}\right)\right\}, \qquad D > 0, \tag{4.16a}$$

$$D^{\parallel} = \frac{\sigma_{\mathrm{x,z}}^2}{\sigma_{\mathrm{x}}^2}(= \sigma_{\mathrm{x}}^2 V^2) = \min\left\{\sigma_{\mathrm{x}}^2\,,\ \frac{D^2}{b^2\sigma_{\mathrm{x}}^2}\right\}, \qquad D > 0.$$

▲

*Proof.* We first note that for any $D \ge b\sigma_{\mathrm{x}}^2$, the choice $V = 1$ and $\sigma_{\mathrm{u}}^2 = \frac{1}{a}(D - b\sigma_{\mathrm{x}}^2)$ yields $D_{a,b}(\mathrm{x},\mathrm{x}+\mathrm{z}) = b\sigma_{\mathrm{x}}^2$ and $I(\mathrm{x};\mathrm{y}) = 0$. Thus, $R_{a,b}(D) = 0$, $\forall D \ge b\sigma_{\mathrm{x}}^2$. Secondly, the mutual information between $\mathrm{x}$ and $(\mathrm{x}+\mathrm{z})$ is given by

$$I(\mathrm{x};\mathrm{x}+\mathrm{z}) = h(\mathrm{x}+\mathrm{z}) - h(\mathrm{x}+\mathrm{z}\,|\,\mathrm{x}) = h(\mathrm{x}+\mathrm{z}) - h(-V\,\mathrm{x}+\mathrm{u}\,|\,\mathrm{x}) = h(\mathrm{x}+\mathrm{z}) - h(\mathrm{u}\,|\,\mathrm{x}) \tag{4.17}$$

where Properties 2.1 and 2.6 have been used (see Section 2.3). In view of Lemma 4.1, the optimal $\mathrm{z}$ must be Gaussian. It thus follows that the optimal $\mathrm{u}$ must be jointly Gaussian with $\mathrm{x}$. The fact that $\mathrm{u}$ is Gaussian, together with (4.11), implies that $\mathrm{u}$ is independent of $\mathrm{x}$. Thus, from (4.17),[2]

$$\begin{aligned} I(\mathrm{x};\mathrm{y}) = h(\mathrm{x}+\mathrm{z}) - h(\mathrm{u}) &= \frac{1}{2}\ln(2\pi\,\mathrm{e}\,\sigma^2_{\mathrm{x}+\mathrm{z}}) - \frac{1}{2}\ln(2\pi\,\mathrm{e}\,\sigma^2_{\mathrm{u}}) \\ &= \frac{1}{2}\ln([1-V]^2\sigma^2_{\mathrm{x}} + \sigma^2_{\mathrm{u}}) - \frac{1}{2}\ln(\sigma^2_{\mathrm{u}}). \end{aligned} \tag{4.18}$$

For any given and fixed value of $D > 0$, the noise variance $\sigma^2_{\mathrm{u}}$ can be expressed in terms of $V$, as follows

$$\sigma^2_{\mathrm{u}} = \frac{D - bV^2\sigma^2_{\mathrm{x}}}{a}. \tag{4.19}$$

Substituting this into (4.18) one obtains

$$I(\mathrm{x};\mathrm{y}) = \frac{1}{2}\ln\left([1-V]^2\sigma^2_{\mathrm{x}} + \frac{D - bV^2\sigma^2_{\mathrm{x}}}{a}\right) - \frac{1}{2}\ln\left(\frac{D - bV^2\sigma^2_{\mathrm{x}}}{a}\right). \tag{4.20}$$

The value of $V$ that minimizes (4.20) needs to satisfy

$$\begin{aligned} 0 = \frac{\partial I(\mathrm{x};\mathrm{y})}{\partial V} &= \frac{-\sigma^2_{\mathrm{x}}[1-V] - \frac{b}{a}\sigma^2_{\mathrm{x}}V}{[1-V]^2\sigma^2_{\mathrm{x}} + \frac{D-bV^2\sigma^2_{\mathrm{x}}}{a}} - \frac{-\frac{b}{a}\sigma^2_{\mathrm{x}}V}{\frac{D-bV^2\sigma^2_{\mathrm{x}}}{a}} \iff \\ 0 &= \left(-\sigma^2_{\mathrm{x}}[1-V] - \frac{b}{a}\sigma^2_{\mathrm{x}}V\right)\left(D - bV^2\sigma^2_{\mathrm{x}}\right) + \frac{b}{a}\sigma^2_{\mathrm{x}}V\left(a[1-V]^2\sigma^2_{\mathrm{x}} + D - bV^2\sigma^2_{\mathrm{x}}\right) \\ &= -\sigma^2_{\mathrm{x}}[1-V]\left(D - bV^2\sigma^2_{\mathrm{x}}\right) + b\sigma^4_{\mathrm{x}}V[1-V]^2 = [1-V]\left(-D + bV^2\sigma^2_{\mathrm{x}} + b\sigma^2_{\mathrm{x}}[1-V]V\right) \\ &= [1-V]\left(b\sigma^2_{\mathrm{x}}V - D\right) \iff \\ V &= \begin{cases} 1 & , \text{ in any case, or} \\ \frac{D}{b\sigma^2_{\mathrm{x}}} & , \text{ if } \frac{D}{b} \le \sigma^2_{\mathrm{x}}. \end{cases} \end{aligned} \tag{4.21}$$

But for any given $D$, the right hand side of (4.19) must be non-negative. Thus, (4.21) becomes

$$V = \min\left\{\frac{D}{b\sigma^2_{\mathrm{x}}}, 1\right\} \tag{4.22}$$

Substituting (4.22) into (4.13) yields (4.16). Finally, substitution of (4.22) into (4.20) yields (4.15). This completes the proof. □

The rate-distortion function characterized above is achievable only when the source is a scalar memoryless process, by encoding (infinitely) long sequences of it, see Section 4.8. In this case, each source element can be taken as a different realization of a single scalar random variable.

[2] Notice that $\sigma^2_{\mathrm{u}}$ is non-zero; otherwise, unless $\mathrm{z} = -\mathrm{x}$, the mutual information between $\mathrm{x}$ and $\mathrm{z}+\mathrm{z}$ would be unbounded.

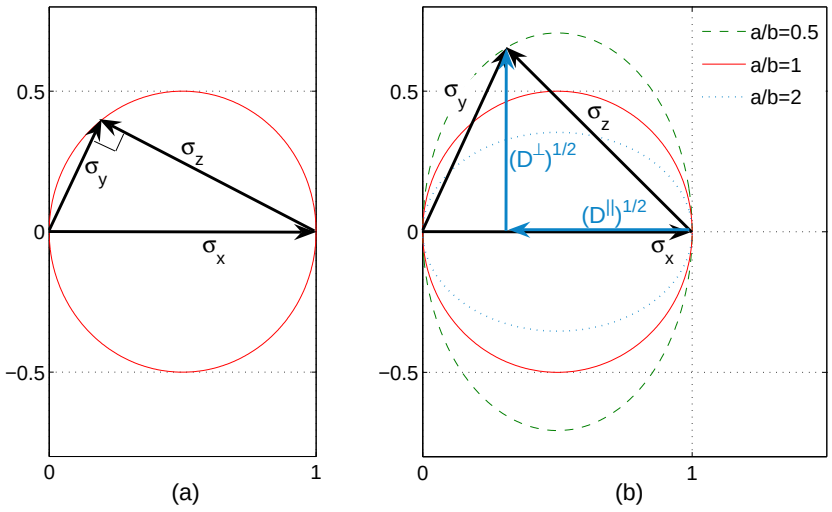



**Figure 4.1:** Geometric locus of the realizations of the rate distortion for a scalar Gaussian source $\mathrm{x}$ and (a): MSE as the distortion metric; (b) WCMSE with weights $a$, $b$ as the distortion metric, according to (4.25). In all plots, $\sigma_{\mathrm{x}}^2 = 1$ and $b = 1$.

### 4.3.1 Geometrical Interpretation

For a Gaussian, zero-mean, scalar random source $\mathrm{x}$, and using MSE as the distortion metric, any realization of Shannon's $R(D)$ is always such that the reconstruction error $\mathrm{z}$ is Gaussian and independent of the reconstruction random variable $\mathrm{y}$. [3] This can be seen as a consequence of the well known fact that, minimum MSE filtering (in this case, scaling), applied to a noisy signal, leaves a reconstruction error which is uncorrelated to the noisy signal, so that

$$\sigma_{\mathrm{x}}^2 = \sigma_{\mathrm{y}}^2 + \sigma_{\mathrm{z}}^2. \tag{4.23}$$

A geometrical interpretation of this fact is shown in Fig. 4.1-(a). In this figure, the vectors labeled with the lengths $\sigma_{\mathrm{x}}$, $\sigma_{\mathrm{y}}$ and $\sigma_{\mathrm{z}}$, represent, respectively, the source, the reconstruction, and the reconstruction error. The squared norms of these vectors are precisely the variances of the variables they represent. Since $\mathrm{y} = \mathrm{x} + \mathrm{z}$, and in view of (4.23), their respective vectors form a right-angled triangle, with $\mathrm{x}$ being the hypotenuse. Therefore, for all possible values of the distortion $\sigma_{\mathrm{z}}^2$, the corresponding outputs describe

[3] In a realization of the quadratic Gaussian RDF, the reconstruction error must be jointly Gaussian with the output (reconstructed) signal. Thus, if the latter were not independent from the output signal, then a Wiener filter applied to the output signal would reduce the MSE, whilst preserving the mutual information. The resulting reconstruction error from a Wiener filter is uncorrelated to the output of the filter, which, for jointly Gaussian signals, implies independence.

a semi-circle, such as the one shown in Fig. 4.1-(a). In other words, every point on the circle represents a point on the $R(D)$ curve of $\mathrm{x}$. This type of plot gives a clear idea of the statistical relationship between source and reconstruction error. Thus, for example, as the distortion becomes smaller (smaller $\sigma_{\mathrm{z}}^2$), the right-angle vertex of the triangle slides to the right along the circle. This implies that the reconstruction error $\mathbf{z}$ not only becomes smaller, but also becomes more orthogonal with respect to the source. In these cases, most of the distortion will be comprised of additive noise, uncorrelated to the source. Conversely, for large distortions the right-angle vertex moves to the left, and the reconstruction error increasingly resembles the negative of the source. This can be seen as strong source attenuation (linear distortion) plus a small amount of source-uncorrelated error. In all cases, the circle shown in Fig. 4.1-(a) determines the balance between source-uncorrelated and source-parallel errors for all possible realizations of $R(D)$.

The situation with the WCMSE rate-distortion function is different to the one just described. As will be shown below, the geometric locus of all pairs of possible values $(D^{\perp}, D^{\parallel})$, stemming from the realization of $R_{a,b}(D)$ for a Gaussian scalar source $\mathrm{x}$, is a family of ellipses in $\mathbb{R}^2$. To see this, recall, from the definition of the WCMSE (see (1.6)), that

$$aD^{\perp} + bD^{\parallel} = D_{a,b}(\mathrm{x}, \mathrm{y}). \tag{4.24}$$

On the other hand (4.16), implies that, when $R_{a,b}(D)$ is realized, $D_{a,b} = b\sigma_{\mathrm{x}}^2\sqrt{D^{\parallel}}$. Substituting the latter into (4.24) we obtain, after some algebra, that

$$\left(\sqrt{D^{\parallel}} - \frac{\sigma_{\mathrm{x}}}{2}\right)^2 + \frac{a}{b}D^{\perp} = \frac{\sigma_{\mathrm{x}}^2}{4}. \tag{4.25}$$

This equation describes a family of ellipses whose vertical-diameter/horizontal-diameter ratio is given by $\sqrt{b/a}$. Three of these ellipses are illustrated in Fig. 4.1-(b). When $a = b = 1$, the ellipse obtained is a circle (solid line in Fig. 4.1-(b)), and WCMSE equals standard MSE. When $a < b$, source-uncorrelated errors are less important that source-parallel errors, and thus the realizations of $R_{a,b}(D)$ lie on ellipses whose vertical axis (perpendicular to $\mathrm{x}$) are larger than their horizontal axis. The opposite situation occurs when $a > b$. It is also interesting to see that when $a \neq b$, and if $b/a < \infty$, the reconstruction error that realizes $R_{a,b}(D)$ is neither orthogonal to the source nor to the reconstruction. In the limit situation in which $a$ is bounded and $b \to \infty$, the ellipse degenerates into two parallel vertical lines, the left line at $-\infty$, the right line passing by the point of the $\sigma_{\mathrm{x}}$ vector. For this extreme case, all the realizations of $R_{a,b}(D)$ are such that the reconstruction error is completely orthogonal to the source.

The plot of $R_{a,b}(D)$ for a fixed value of $a$ and several values of the parameter $b$ is shown in Fig. 4.2. For all these curves, $a$=1. As expected from (4.15), the distortion level at which the rate becomes zero (and the source is reconstructed simply by its mean value) equals the value of $b$. In the limit as $b \to \infty$,

the plot of $R_{a,b}(D)$ approaches zero only asymptotically as $D \to \infty$.

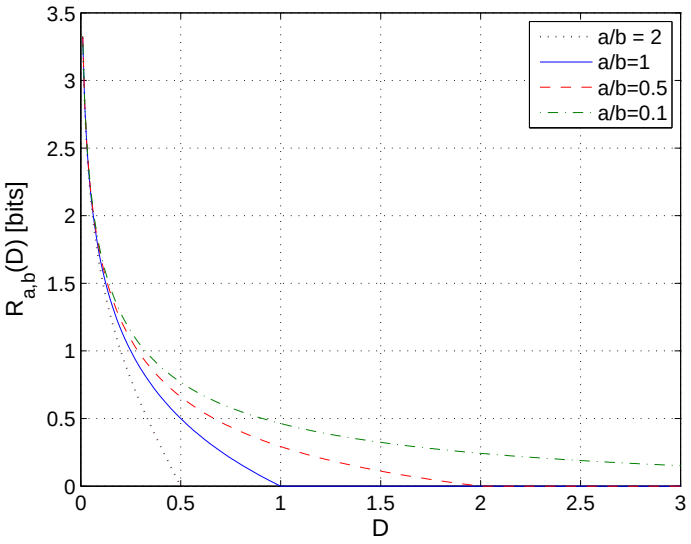



**Figure 4.2:** $R_{a,b}$ for a unit-variance Gaussian scalar source. In all plots, $a$=1.

### 4.3.2 Convexity of $R_{a,b}(D)$

It will be shown below that the WCMSE rate-distortion function characterized above *is not convex for some choices of weights* $a$, $b$. For this purpose, we take the first derivative of $R_{a,b}(D)$. Specifically, differentiation of (4.15) yields

$$\frac{dR_{a,b}(D)}{d\,D} = -\frac{1}{2} \cdot \frac{a\sigma_{\mathrm{x}}^2}{a\sigma_{\mathrm{x}}^2 D + [1-\frac{a}{b}]D^2}, \quad D \leq b\sigma_{\mathrm{x}}^2.$$

Differentiating again,

$$\frac{d^2R_{a,b}(D)}{d\,D^2} = \frac{a\sigma_{\mathrm{x}}^2}{2}\left(a\sigma_{\mathrm{x}}^2 D + [1-\tfrac{a}{b}]D^2\right)^{-2}\left(a\sigma_{\mathrm{x}}^2 + 2[1-\tfrac{a}{b}]D\right), \quad D \leq b\sigma_{\mathrm{x}}^2.$$

Thus, for distortions within the interval $(0, b\sigma_{\mathrm{x}}^2]$,

$$\frac{d^2R_{a,b}(D)}{d\,D^2} \geq 0 \iff 0 \leq a\sigma_{\mathrm{x}}^2 + 2[1-\tfrac{a}{b}]D \tag{4.26}$$

If $a \geq b$, then $R_{a,b}(D)$ is convex. Otherwise, the condition in (4.26) becomes

$$\frac{d^2R_{a,b}(D)}{d\,D^2} \geq 0 \iff -a\sigma_{\mathrm{x}}^2 \leq 2[1-\tfrac{a}{b}]D \iff \frac{a\sigma_{\mathrm{x}}^2}{2[\frac{a}{b}-1]} \geq D. \tag{4.27}$$

As can be seen from (4.27), if $b < a$, then $R_{a,b}(D)$ is convex only over the union of intervals $(0\,,\, a\sigma_{\mathrm{x}}^2/(2[\frac{a}{b}-1])] \cup (b\sigma_{\mathrm{x}}^2, \infty)$. Therefore, $R_{a,b}(D)$ is convex over the positive real line if and only if

$$\frac{a\sigma_{\mathrm{x}}^2}{2[\frac{a}{b}-1]} \geq b\sigma_{\mathrm{x}}^2 \iff \frac{a}{2[a-b]} \geq 1 \iff \frac{a}{b} \leq 2 \tag{4.28}$$

For future reference, we summarize the above result in the form of the following lemma.

**Lemma 4.3.** *For a zero mean, Gaussian and scalar random source, $R_{a,b}(D)$ is convex if and only if $a < 2b$.* ▲

The fact that $R_{a,b}(D)$ is non-convex for certain choices of $a$, $b$, may seem, at first, surprising. After all, it is well known that any rate-distortion function originating from a single-letter fidelity criterion is convex, as shown in [6, Theorem 2.4.1]. However, the proof of Theorem 2.4.1 in [6] relies upon the fact that the distortion is linear in the joint probability distribution of source and reconstruction. Thus, the possible non-convexity of $R_{a,b}(D)$ does not contradict the latter theorem since, as shown in Section 4.2.1, $D_{a,b}(\mathrm{x},\mathrm{y})$ is not linear in $f_{\mathrm{y}\,|\,\mathrm{x}}(\cdot|\cdot)$.

The dependence of $R_{a,b}(D)$ on the ratio $a/b$ is illustrated in Fig. 4.3. In this figure, four plots of $R_{a,b}(D)$ are displayed for a unit-variance Gaussian source and a fixed WCMSE weight value $b = 1$. Each plot corresponds to a different value of the ratio $a/b$. Notice from this figure that, for $a/b = 2$, $R_{a,b}(D)$ is still convex, although its plot is almost a straight line for distortions close to $b\sigma_{\mathrm{x}}^2$. On the other hand, as predicted by (4.28), the choice $a/b = 10$ yields a non-convex $R_{a,b}(D)$, as can be seen in Fig. 4.3 ($-\cdot-$ line plot).

Because convexity of $R_{a,b}(D)$ is required for most of the results to be obtained in the sequel, we shall restrict our analysis, from here on, to weights $a$, $b$ such that $b \geq a/2$. In doing so, we will leave aside situations in which the cost of source-parallel error is less than half the cost of source-uncorrelated error. Although such situations could arise in practice, it can be argued that the condition $b > a/2$ holds for many cases of interest. First, it holds whenever it is important to achieve, or closely approximate, a given signal transfer function, as described in Section 1.1. Second, the fact that the use of dither in audio and image quantization yields noise that is perceived as more acceptable by human listeners/observers [156–158], suggests that the perceptual cost of source-uncorrelated noise in image and audio encoding is, in general, less than that of source-parallel noise. The reader may argue against the validity of latter statement since, for instance, simple forms of source parallel distortion, such as a small delay (in audio) or pixel shift (in images), are barely objectionable by a human listener or observer. Nevertheless, loosely speaking, from a data compression viewpoint there is little to gain from such forms of distortion. (Recall that altering the phase of a random source has no effect on its entropy). In other words, it can be conjectured that source-parallel distortion can provide data compression only when it takes the form

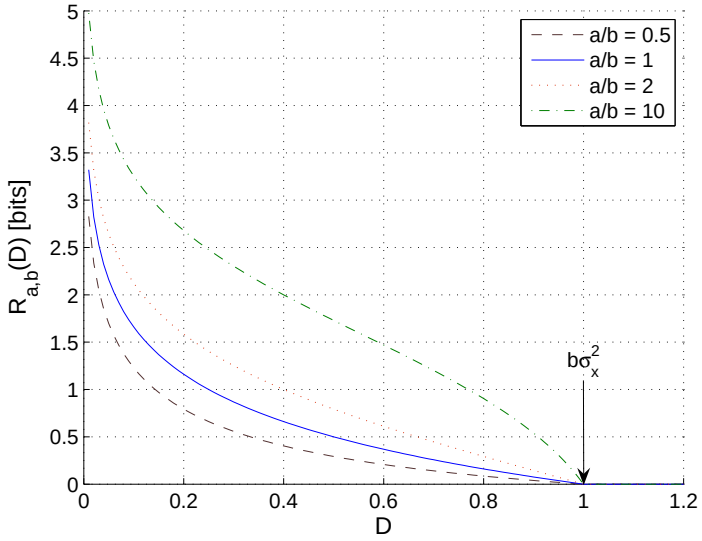



**Figure 4.3:** $R_{a,b}(D)$ for a unit-variance Gaussian scalar source. In all the plots $b = 1$.

of attenuation of the power of the source (in the presence of noise), or when it involves discarding part of its information content. The perceptual impact of such "compressive" forms of source-parallel distortion is likely to be at least comparable to the perceptual impact of source-uncorrelated distortion, hence satisfying $b > a/2$. Of course, the degree of validity of the latter assumption will ultimately depend on each particular application. The validity of the latter argument in lossy image compression is supported by the sequence of images shown in the example in Section 4.7, on page 143.

## 4.4 WCMSE RDF For Gaussian Vector Sources

In this section we derive the WCMSE-rate distortion function for Gaussian vectors. Before proceeding, we need to establish some preliminary results.

### 4.4.1 Preliminary Results

The following Lemma will be useful in subsequent derivations.

**Lemma 4.4.** *Let $\mathbf{v}_1, \mathbf{v}_2$ be two mutually independent random vectors. Then, the following holds:*

$$\bar{I}(\mathbf{v}_1, \mathbf{v}_2\,;\, \mathbf{w}_1, \mathbf{w}_2) \geq \bar{I}(\mathbf{v}_1; \mathbf{w}_1) + \bar{I}(\mathbf{v}_2; \mathbf{w}_2) \tag{4.29}$$

*Equality holds if and only if the random vectors* $\mathbf{n}_1 \triangleq \mathbf{w}_1 - \mathbf{v}_1$ *and* $\mathbf{n}_2 \triangleq \mathbf{w}_2 - \mathbf{v}_2$ *are such that*

$$f_{\mathbf{n}_1,\mathbf{n}_2,\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{v}_1,\boldsymbol{v}_2) = f_{\mathbf{n}_1,\mathbf{v}_1}(\boldsymbol{n}_1,\boldsymbol{v}_1) f_{\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_2,\boldsymbol{v}_2) \tag{4.30}$$

*almost everywhere.* ▲

*Proof.* We proceed by parts. We first prove (4.29), and then the necessity and sufficiency of (4.30) for achieving equality in (4.29).

1. ($\geq$): We have that

$$\begin{aligned}
\bar{I}(\mathbf{v}_1,\mathbf{v}_2\,;\,\mathbf{w}_1,\mathbf{w}_2) &= \bar{h}(\mathbf{v}_1,\mathbf{v}_2) - \bar{h}(\mathbf{v}_1,\mathbf{v}_2|\mathbf{w}_1,\mathbf{w}_2) \\
&= \bar{h}(\mathbf{v}_1,\mathbf{v}_2) - \bar{h}(\mathbf{n}_1,\mathbf{n}_2|\mathbf{w}_1,\mathbf{w}_2) \\
&\overset{(a)}{\geq} \bar{h}(\mathbf{v}_1,\mathbf{v}_2) - \bar{h}(\mathbf{n}_1|\mathbf{w}_1,\mathbf{w}_2) - \bar{h}(\mathbf{n}_2|\mathbf{w}_1,\mathbf{w}_2) \\
&\overset{(b)}{\geq} \bar{h}(\mathbf{v}_1,\mathbf{v}_2) - \bar{h}(\mathbf{n}_1|\mathbf{w}_1) - \bar{h}(\mathbf{n}_2|\mathbf{w}_2) \\
&= \bar{h}(\mathbf{v}_1) + \bar{h}(\mathbf{v}_2) - \bar{h}(\mathbf{n}_1|\mathbf{w}_1) - \bar{h}(\mathbf{n}_2|\mathbf{w}_2). \\
&= \bar{h}(\mathbf{v}_1) + \bar{h}(\mathbf{v}_2) - \bar{h}(\mathbf{v}_1|\mathbf{w}_1) - \bar{h}(\mathbf{v}_2|\mathbf{w}_2). \\
&= \bar{I}(\mathbf{v}_1;\mathbf{w}_1) + \bar{I}(\mathbf{v}_2;\mathbf{w}_2).
\end{aligned}$$

   Equality is achieved in $(a)$ if and only if the following Markov chain holds

$$\mathbf{n}_1 \longleftrightarrow \{\mathbf{w}_1,\mathbf{w}_2\} \longleftrightarrow \mathbf{n}_2. \tag{4.31}$$

   (See Property 2.3 and Definition 2.18 in Section 2.3.) Similarly, equality holds in $(b)$ if and only if $\mathbf{n}_1, \mathbf{n}_2, \mathbf{w}_1$ and $\mathbf{w}_2$ satisfy the following Markov chains

$$\mathbf{w}_2 \longleftrightarrow \mathbf{w}_1 \longleftrightarrow \mathbf{n}_1, \tag{4.32}$$

$$\mathbf{w}_1 \longleftrightarrow \mathbf{w}_2 \longleftrightarrow \mathbf{n}_2. \tag{4.33}$$

   This follows directly from Property 2.2 and Definition 2.18, see Section 2.3. This establishes (4.29).

2. ($\Rightarrow$): The Markov chain in (4.31) is equivalent to

$$f_{\mathbf{n}_1,\mathbf{n}_2|\mathbf{w}_1,\mathbf{w}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2|\boldsymbol{w}_1,\boldsymbol{w}_2) = f_{\mathbf{n}_1|\mathbf{w}_1,\mathbf{w}_2}(\boldsymbol{n}_1|\boldsymbol{w}_1,\boldsymbol{w}_2) f_{\mathbf{n}_2|\mathbf{w}_1,\mathbf{w}_2}(\boldsymbol{n}_2|\boldsymbol{w}_1,\boldsymbol{w}_2), \tag{4.34}$$

$$\forall \boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{w}_1,\boldsymbol{w}_2.$$

Similarly, the Markov chain in (4.32) is equivalent to

$$\begin{aligned} & f_{\mathbf{n}_1,\mathbf{w}_2|\mathbf{w}_1}(\boldsymbol{n}_1,\boldsymbol{w}_2|\boldsymbol{w}_1) = f_{\mathbf{n}_1|\mathbf{w}_1}(\boldsymbol{n}_1|\boldsymbol{w}_1) f_{\mathbf{w}_2|\mathbf{w}_1}(\boldsymbol{w}_2|\boldsymbol{w}_1), \quad \forall \boldsymbol{n}_1,\boldsymbol{w}_1,\boldsymbol{w}_2 \\ \iff & f_{\mathbf{n}_1|\mathbf{w}_2,\mathbf{w}_1}(\boldsymbol{n}_1|\boldsymbol{w}_2,\boldsymbol{w}_1) = f_{\mathbf{n}_1|\mathbf{w}_1}(\boldsymbol{n}_1|\boldsymbol{w}_1), \quad \forall \boldsymbol{w}_1,\boldsymbol{w}_2 : f_{\mathbf{w}_2|\mathbf{w}_1}(\boldsymbol{w}_2|\boldsymbol{w}_1) > 0, \end{aligned} \tag{4.35}$$

whilst the Markov chain (4.33) is equivalent to

$$\begin{aligned} & f_{\mathbf{n}_2,\mathbf{w}_1|\mathbf{w}_2}(\boldsymbol{n}_2,\boldsymbol{w}_1|\boldsymbol{w}_2) = f_{\mathbf{n}_2|\mathbf{w}_2}(\boldsymbol{n}_2|\boldsymbol{w}_2) f_{\mathbf{w}_1|\mathbf{w}_2}(\boldsymbol{w}_1|\boldsymbol{w}_2), \quad \forall \boldsymbol{n}_2,\boldsymbol{w}_1,\boldsymbol{w}_2 \\ \iff & f_{\mathbf{n}_2|\mathbf{w}_2,\mathbf{w}_1}(\boldsymbol{n}_2|\boldsymbol{w}_2,\boldsymbol{w}_1) = f_{\mathbf{n}_2|\mathbf{w}_2}(\boldsymbol{n}_2|\boldsymbol{w}_2), \quad \forall \boldsymbol{w}_1,\boldsymbol{w}_2 : f_{\mathbf{w}_1|\mathbf{w}_2}(\boldsymbol{w}_1|\boldsymbol{w}_2) > 0. \end{aligned} \tag{4.36}$$

Substitution of (4.35) and (4.36) into (4.34) yields

$$\begin{aligned} f_{\mathbf{n}_1,\mathbf{n}_2|\mathbf{w}_1,\mathbf{w}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2|\boldsymbol{w}_1,\boldsymbol{w}_2) = f_{\mathbf{n}_1|\mathbf{w}_1}(\boldsymbol{n}_1|\boldsymbol{w}_1) f_{\mathbf{n}_2|\mathbf{w}_2}(\boldsymbol{n}_2|\boldsymbol{w}_2), \\ \forall \boldsymbol{w}_1,\boldsymbol{w}_2 : f_{\mathbf{w}_1,\mathbf{w}_2}(\boldsymbol{w}_1,\boldsymbol{w}_2) > 0. \end{aligned} \tag{4.37}$$

On the other hand, since $\mathbf{w}_1 = \mathbf{v}_1 + \mathbf{n}_1$ and $\mathbf{w}_2 = \mathbf{v}_2 + \mathbf{n}_2$, we have that

$$f_{\mathbf{n}_1,\mathbf{n}_2|\mathbf{w}_1,\mathbf{w}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2|\boldsymbol{w}_1,\boldsymbol{w}_2) = f_{\mathbf{n}_1,\mathbf{n}_2|\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2|\boldsymbol{w}_1-\boldsymbol{n}_1,\boldsymbol{w}_2-\boldsymbol{n}_2), \tag{4.38a}$$

$$f_{\mathbf{n}_1|\mathbf{w}_1}(\boldsymbol{n}_1|\boldsymbol{w}_1) = f_{\mathbf{n}_1|\mathbf{v}_1}(\boldsymbol{n}_1|\boldsymbol{w}_1-\boldsymbol{n}_1), \text{ and} \tag{4.38b}$$

$$f_{\mathbf{n}_2|\mathbf{w}_2}(\boldsymbol{n}_2|\boldsymbol{w}_2) = f_{\mathbf{n}_2|\mathbf{v}_2}(\boldsymbol{n}_2|\boldsymbol{w}_2-\boldsymbol{n}_2), \tag{4.38c}$$

for all $\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{w}_1,\boldsymbol{w}_2$. Substitution of (4.38) into (4.37) yields

$$\begin{aligned} f_{\mathbf{n}_1,\mathbf{n}_2|\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2|\boldsymbol{w}_1-\boldsymbol{n}_1,\boldsymbol{w}_2-\boldsymbol{n}_2) &= f_{\mathbf{n}_1|\mathbf{v}_1}(\boldsymbol{n}_1|\boldsymbol{w}_1-\boldsymbol{n}_1) f_{\mathbf{n}_2|\mathbf{v}_2}(\boldsymbol{n}_2|\boldsymbol{w}_2-\boldsymbol{n}_2), \\ & \qquad \forall \boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{w}_1,\boldsymbol{w}_2 : f_{\mathbf{w}_1,\mathbf{w}_2}(\boldsymbol{w}_1,\boldsymbol{w}_2) > 0 \\ \iff f_{\mathbf{n}_1,\mathbf{n}_2|\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2|\boldsymbol{v}_1,\boldsymbol{v}_2) &= f_{\mathbf{n}_1|\mathbf{v}_1}(\boldsymbol{n}_1|\boldsymbol{v}_1) f_{\mathbf{n}_2|\mathbf{v}_2}(\boldsymbol{n}_2|\boldsymbol{v}_2), \\ & \qquad \forall \boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{v}_1,\boldsymbol{v}_2 : f_{\mathbf{w}_1,\mathbf{w}_2}(\boldsymbol{v}_1+\boldsymbol{n}_1,\boldsymbol{v}_2+\boldsymbol{n}_2) > 0. \\ \implies f_{\mathbf{n}_1,\mathbf{n}_2,\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{v}_1,\boldsymbol{v}_2) &= f_{\mathbf{n}_1,\mathbf{v}_1}(\boldsymbol{n}_1,\boldsymbol{v}_1) f_{\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_2,\boldsymbol{v}_2), \\ & \qquad \forall \boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{v}_1,\boldsymbol{v}_2 : f_{\mathbf{n}_1,\mathbf{n}_2,\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{v}_1,\boldsymbol{v}_2) > 0. \end{aligned} \tag{4.39}$$

The last implication in the above follows on multiplying both sides of the preceeding equation by $f_{\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{v}_1,\boldsymbol{v}_2) = f_{\mathbf{v}_1}(\boldsymbol{v}_1) f_{\mathbf{v}_2}(\boldsymbol{v}_2)$ (recall that $\mathbf{v}_1$ and $\mathbf{v}_2$ are independent), and from the fact that

$$f_{\mathbf{n}_1,\mathbf{n}_2,\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{v}_1,\boldsymbol{v}_2) > 0 \implies f_{\mathbf{w}_1,\mathbf{w}_2}(\boldsymbol{v}_1+\boldsymbol{n}_1,\boldsymbol{v}_2+\boldsymbol{n}_2) > 0, \qquad \forall \boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{v}_1,\boldsymbol{v}_2.$$

It is only left to demonstrate that (4.39) implies (4.30). In view of (4.39), if (4.30) does not hold, then there must exist a set of vectors $\mathbb{P}$, having non-zero measure, such that, for all $\{\boldsymbol{n}_1',\boldsymbol{n}_2',\boldsymbol{v}_1',\boldsymbol{v}_2'\} \in \mathbb{P}$,

$$f_{\mathbf{n}_1,\mathbf{n}_2,\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{n}_1',\boldsymbol{n}_2',\boldsymbol{v}_1',\boldsymbol{v}_2') = 0 \tag{4.40}$$

while

$$f_{\mathbf{n}_1,\mathbf{v}_1}(\boldsymbol{n}_1',\boldsymbol{v}_1')f_{\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_2',\boldsymbol{v}_2')>0. \tag{4.41}$$

Clearly, (4.41) implies that $f_{\mathbf{n}_1,\mathbf{v}_1}(\boldsymbol{n}_1',\boldsymbol{v}_1')>0$ and that $f_{\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_2',\boldsymbol{v}_2')>0$. Since

$$f_{\mathbf{n}_1,\mathbf{v}_1}(\boldsymbol{n}_1',\boldsymbol{v}_1')=\iint f_{\mathbf{n}_1,\mathbf{n}_2,\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{n}_1',\boldsymbol{n}_2,\boldsymbol{v}_1',\boldsymbol{v}_2)d\boldsymbol{n}_2d\boldsymbol{v}_2, \quad \text{and} \tag{4.42}$$

$$f_{\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_2',\boldsymbol{v}_2')=\iint f_{\mathbf{n}_1,\mathbf{n}_2,\mathbf{v}_1,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2',\boldsymbol{v}_1,\boldsymbol{v}_2')d\boldsymbol{n}_1d\boldsymbol{v}_1, \tag{4.43}$$

it follows that, for all $\{\boldsymbol{n}_1',\boldsymbol{n}_2',\boldsymbol{v}_1',\boldsymbol{v}_2'\}\in\mathbb{P}$, the sets

$$\mathbb{S}_1(\boldsymbol{n}_2',\boldsymbol{v}_2')\triangleq\{\{\boldsymbol{n}_1,\boldsymbol{v}_1\}\ :\ f_{\mathbf{n}_1,\mathbf{v}_1,\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{n}_2',\boldsymbol{v}_1,\boldsymbol{v}_2')>0\} \tag{4.44}$$

$$\mathbb{S}_2(\boldsymbol{n}_1',\boldsymbol{v}_1')\triangleq\{\{\boldsymbol{n}_2,\boldsymbol{v}_2\}\ :\ f_{\mathbf{n}_1,\mathbf{v}_1,\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_1',\boldsymbol{n}_2,\boldsymbol{v}_1',\boldsymbol{v}_2)>0\} \tag{4.45}$$

have non-zero measure. On the other hand, dividing both sides of (4.39) by either $f_{\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_2',\boldsymbol{v}_2')$ or by $f_{\mathbf{n}_1,\mathbf{v}_1}(\boldsymbol{n}_1',\boldsymbol{v}_1')$, we obtain, for each case,

$$f_{\mathbf{n}_1,\mathbf{v}_1|\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{v}_1|\boldsymbol{n}_2',\boldsymbol{v}_2')=f_{\mathbf{n}_1,\mathbf{v}_1}(\boldsymbol{n}_1,\boldsymbol{v}_1),\quad \forall\{\boldsymbol{n}_1,\boldsymbol{v}_1\}\in\mathbb{S}_1(\boldsymbol{n}_2',\boldsymbol{v}_2'), \tag{4.46}$$

$$f_{\mathbf{n}_2,\mathbf{v}_2|\mathbf{n}_1,\mathbf{v}_1}(\boldsymbol{n}_2,\boldsymbol{v}_2|\boldsymbol{n}_1',\boldsymbol{v}_1')=f_{\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_2,\boldsymbol{v}_2),\quad \forall\{\boldsymbol{n}_2,\boldsymbol{v}_2\}\in\mathbb{S}_2(\boldsymbol{n}_1',\boldsymbol{v}_1'). \tag{4.47}$$

From (4.46) and (4.44), we have:

$$1=\iint\limits_{\{\boldsymbol{n}_1,\boldsymbol{v}_1\}\in\mathbb{S}_1(\boldsymbol{n}_2',\boldsymbol{v}_2')} f_{\mathbf{n}_1,\mathbf{v}_1|\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_1,\boldsymbol{v}_1|\boldsymbol{n}_2',\boldsymbol{v}_2')d\boldsymbol{n}_1d\boldsymbol{v}_1=\iint\limits_{\{\boldsymbol{n}_1,\boldsymbol{v}_1\}\in\mathbb{S}_1(\boldsymbol{n}_2',\boldsymbol{v}_2')} f_{\mathbf{n}_1,\mathbf{v}_1}(\boldsymbol{n}_1,\boldsymbol{v}_1)d\boldsymbol{n}_1d\boldsymbol{v}_1.$$

This implies that $f_{\mathbf{n}_1,\mathbf{v}_1}(\boldsymbol{n}_1,\boldsymbol{v}_1)=0$ almost everywhere outside $\mathbb{S}_1(\boldsymbol{n}_2',\boldsymbol{v}_2')$. A similar analysis yields that $f_{\mathbf{n}_2,\mathbf{v}_2}(\boldsymbol{n}_2,\boldsymbol{v}_2)=0$ almost everywhere outside $\mathbb{S}_2(\boldsymbol{n}_1',\boldsymbol{v}_1')$. It then follows that

$$\int\limits_{\mathbb{P}} d\boldsymbol{n}_1'd\boldsymbol{n}_2'd\boldsymbol{v}_1'd\boldsymbol{v}_2'=\int\limits_{\boldsymbol{n}_1',\boldsymbol{v}_1'}\left(\int\limits_{\{\boldsymbol{n}_2',\boldsymbol{v}_2'\}\notin\mathbb{S}_2(\boldsymbol{n}_1',\boldsymbol{v}_1')} d\boldsymbol{n}_2'd\boldsymbol{v}_2'\right)d\boldsymbol{n}_1'd\boldsymbol{v}_1'=0, \tag{4.48}$$

and thus $\mathbb{P}$ has zero measure. This proves that achieving equality in (4.29) implies (4.30).

3. ($\Leftarrow$). Here it will be shown that (4.30) implies that equality holds in (4.29). We have that

$$\begin{aligned}
\bar{I}(\mathbf{v}_1,\mathbf{v}_2;\mathbf{y}_1,\mathbf{y}_2)&=\bar{h}(\mathbf{y}_1,\mathbf{y}_2)-\bar{h}(\mathbf{y}_1,\mathbf{y}_2|\mathbf{v}_1,\mathbf{v}_2)\\
&=\bar{h}(\mathbf{y}_1,\mathbf{y}_2)-\bar{h}(\mathbf{n}_1,\mathbf{n}_2|\mathbf{v}_1,\mathbf{v}_2)\\
&=\bar{h}(\mathbf{y}_1,\mathbf{y}_2)-\bar{h}(\mathbf{n}_1|\mathbf{v}_1,\mathbf{v}_2)-\bar{h}(\mathbf{n}_2|\mathbf{v}_1,\mathbf{v}_2,\mathbf{n}_1)\\
&\overset{(a)}{=}\bar{h}(\mathbf{y}_1,\mathbf{y}_2)-\bar{h}(\mathbf{n}_1|\mathbf{v}_1)-\bar{h}(\mathbf{n}_2|\mathbf{v}_2)\\
&\overset{(b)}{=}\bar{h}(\mathbf{y}_1)+\bar{h}(\mathbf{y}_2)-\bar{h}(\mathbf{n}_1|\mathbf{v}_1)-\bar{h}(\mathbf{n}_2|\mathbf{v}_2)\\
&=\bar{I}(\mathbf{v}_1;\mathbf{y}_1)+\bar{I}(\mathbf{v}_2;\mathbf{y}_2)
\end{aligned}$$

where $(a)$ follows directly from (4.30). Also, $(b)$ stems from Property 2.3 (on page 37), and from the fact that (4.30) implies $f_{\mathbf{y}_1,\mathbf{y}_2}(\boldsymbol{y}_1,\boldsymbol{y}_2) = f_{\mathbf{y}_1}(\boldsymbol{y}_1) f_{\mathbf{y}_2}(\boldsymbol{y}_2)$ almost everywhere. This completes the proof.

□

The above lemma allows one to establish the following result:

**Theorem 4.5.** *Let $\{\mathbf{v}_i\}_{i=1}^N$ be mutually independent random vectors (i.e., a vector product source). Let $D\left(\{\mathbf{v}_i\}_{i=1}^N, \{\mathbf{w}_i\}_{i=1}^N\right)$ be a sum distortion metric, i.e., one that satisfies $D\left(\{\mathbf{v}_i\}_{i=1}^N, \{\mathbf{w}_i\}_{i=1}^N\right) = \sum_{i=1}^N D_i\left(\mathbf{v}_i, \mathbf{w}_i\right)$. Denote the RDF of each vector $\mathbf{v}_i$ with respect to the distortion metric $D_i(\cdot,\cdot)$ as $r_i(d)$, and assume that all the functions $r_i(\cdot)$ are convex. Define the scalars $d_i^{max} \triangleq \min\{d : r_i(d) = 0\}$. Let $R(D)$ be the rate-distortion function of $\{\mathbf{v}_i\}_{i=1}^N$ with respect to the distortion metric $D(\cdot,\cdot)$. Then*

$$R(D) = \sum_{i=1}^N r_i(d_i), \tag{4.49}$$

$$D = \sum_{i=1}^N d_i, \tag{4.50}$$

*where distortions $d_i$ are such that* [4]

$$\text{if} \quad r_i'(d_i^{max}) \geq s, \ \text{then} \quad r_i'(d_i) = s, \quad \text{or else,} \tag{4.51a}$$

$$\text{if} \quad r_i'(d_i^{max}) < s, \ \text{then} \quad d_i = d_i^{max}, \tag{4.51b}$$

*for some common slope $s < 0$. Moreover, $R(D)$ is achieved if and only if the distortion random vectors $\mathbf{n}_i \triangleq \mathbf{w}_i - \mathbf{v}_i$, $i = 1, 2, \ldots, N$, are such that*

$$f_{\mathbf{n}_1,\mathbf{n}_2,\ldots,\mathbf{n}_N,\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_N}(\boldsymbol{n}_1,\boldsymbol{n}_2,\ldots,\boldsymbol{n}_N,\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_N) = \prod_{i=1}^N f_{\mathbf{n}_i,\mathbf{v}_i}(\boldsymbol{n}_i,\boldsymbol{v}_i) \tag{4.52}$$

*almost everywhere.* ▲

*Proof.* From Lemma 4.4, we have that

$$\bar{I}\left(\{\mathbf{v}_i\}_{i=1}^N ; \{\mathbf{w}_i\}_{i=1}^N\right) \geq \sum_{i=1}^N \bar{I}\left(\mathbf{v}_i ; \mathbf{w}_i\right) \tag{4.53}$$

with equality if and only if (4.52) holds. On the other hand, the right hand side of (4.53), as well as $D\left(\{\mathbf{v}_i\}_{i=1}^N, \{\mathbf{w}_i\}_{i=1}^N\right)$ depend only on the PDFs $\{f_{\mathbf{n}_i,\mathbf{v}_i}\}_{i=1}^N$. Thus, for any value of

[4] Here the notation $r'(d)$ denotes the derivative of $r(d)$ with respect to $d$. If $r'(d)$ is discontinuous at $d$, then $r'(d)$ denotes the left-derivative of $r(d)$ with respect to $d$.

$D\left(\{\mathbf{v}_i\}_{i=1}^N, \{\mathbf{w}_i\}_{i=1}^N\right)$, $\bar{I}\left(\{\mathbf{v}_i\}_{i=1}^N ; \{\mathbf{w}_i\}_{i=1}^N\right)$ is minimized if and only if (4.52) holds. Therefore, $R(D)$ is achieved if and only if (4.52) holds, and

$$\begin{aligned} R(D) &= \min_{\{\mathbf{w}_i\}_{i=1}^N : D\left(\{\mathbf{v}_i\}_{i=1}^N, \{\mathbf{w}_i\}_{i=1}^N\right) \leq D} \bar{I}\left(\{\mathbf{v}_i\}_{i=1}^N ; \{\mathbf{w}_i\}_{i=1}^N\right) \\ &= \min_{\{\mathbf{w}_i\}_{i=1}^N : D\left(\{\mathbf{v}_i\}_{i=1}^N, \{\mathbf{w}_i\}_{i=1}^N\right) \leq D} \sum_{i=1}^N \bar{I}\left(\mathbf{v}_i ; \mathbf{w}_i\right) \\ &= \min_{\{d_i\}_{i=1}^N : \sum_{i=1}^N d_i \leq D} \sum_{i=1}^N r_i(d_i). \end{aligned} \tag{4.54}$$

The distortions $\{d_i\}_{i=1}^N$ that solve the minimization problem on the right-hand side of (4.54) are such that the Lagrangian

$$\mathscr{L} \triangleq \sum_{i=1}^N r_i(d_i) + s\left(D - \sum_{i=1}^N r_i(d)\right) \tag{4.55}$$

is minimized, where $s$ is a Lagrange multiplier. Hence, the following must hold

$$r_i'(d_i) - s = 0, \quad \forall i = 1, 2, \ldots, N. \tag{4.56}$$

Since the functions $r_i(d)$ are convex, the distortions $d_i$ that satisfy (4.56) are unique. In addition, if for some $i \in \{1, 2, \ldots, N\}$, $r_i'(d_i^{max}) < s$, then $r_i(d)$ has a "corner" at $d = d_i^{max}$, and the slope at this point can be assumed to take any value between $r_i'(d_i^{max})$ and $0$. This, together with (4.56), leads directly to (4.51), completing the proof. □

**Remark 4.2.** *The result in Theorem 4.5 can be seen as an extension of Theorem 2.8.1 and Corollary 2.8.1 in [6] to continuous random variables. However, we believe that Theorem 4.5 improves on [6, Theorem 2.8.1], by actually showing that the $R(D)$-achieving probability assignments are unique, and that* (4.52) *is necessary. The latter claim is indeed present the statement of Theorem 2.8.1 in [6], which, near its end, reads: "[...] Moreover, the conditional probability assignment that yields $R(D_s)$ is the product of the assignments that yield $R_\alpha(D_s^\alpha)$ and $R_\beta(D_s^\beta)$". The latter can be understood as saying that such probability assignment that realizes $R(D)$ is unique and that* (4.52) *is necessary for achieving $R(D)$. However, the proof of Theorem 2.8.1in [6] only shows that, in the notation of this thesis,* (4.52) *(together with* (4.51)*) is* sufficient *for achieving $R(D)$. Thus, Theorem 4.5 improves on [6, Theorem 2.8.1] by actually showing that the $R(D)$-achieving probability assignments are unique, and that* (4.52) *is indeed necessary.* ▲

### 4.4.2 $R_{a,b}(D)$ for Gaussian Vectors

For a vector random source $\mathbf{x} \in \mathbb{R}^N$ reconstructed as $\mathbf{y}$, the reconstruction error is

$$\mathbf{z} \triangleq \mathbf{y} - \mathbf{x}. \tag{4.57}$$

As in the scalar case, $\mathbf{z}$ can always be decomposed into a source-uncorrelated term

$$\mathbf{u} \triangleq \mathbf{z} + \boldsymbol{V}\mathbf{x},$$

and the remainder $-\boldsymbol{V}\mathbf{x}$, which corresponds to the source-parallel term. The *linear distortion* matrix $\boldsymbol{V} \in \mathbb{R}^{N\times N}$ is defined as

$$\boldsymbol{V} \triangleq -\boldsymbol{K}_{\mathbf{z},\mathbf{x}}\boldsymbol{K}_{\mathbf{x}}^{-1}. \tag{4.58}$$

Thus, the reconstruction error can be written as

$$\mathbf{z} = \mathbf{u} - \boldsymbol{V}\mathbf{x} \tag{4.59}$$

where $\mathbf{u}$ is such that $\mathrm{E}\left[\mathbf{u}\mathbf{x}^T\right] = \mathbf{0}$. From this, it follows that the WCMSE for the vector case takes the form

$$D_{a,b}(\mathbf{x},\mathbf{y}) = \frac{1}{N}\,\mathrm{tr}\left\{a\boldsymbol{K}_{\mathbf{u}} + b\boldsymbol{V}\boldsymbol{K}_{\mathbf{x}}\boldsymbol{V}^T\right\}. \tag{4.60}$$

We can now define the WCMSE rate-distortion function for random vectors.

**Definition 4.2.** *The WCMSE-Rate-Distortion Function for a random vector source $\mathbf{x} \in \mathbb{R}^N$ is defined as*

$$R_{a,b}(D) \triangleq \min_{\mathbf{z}:D_{a,b}(\mathbf{x},\mathbf{x}+\mathbf{z})\le D} \frac{1}{N} I(\mathbf{x};\mathbf{x}+\mathbf{z})$$

▲

The following theorem characterizes $R_{a,b}(D)$ for Gaussian vector sources:

**Theorem 4.6** ($R_{a,b}(D)$ for Gaussian Vector Sources)**.** *The rate-distortion function for a Gaussian random vector source with zero mean and covariance matrix $\boldsymbol{K}_{\mathbf{x}}$, with respect to the WCMSE distortion metric with weights $a, b > 0$, where $a \le 2b$, is given by*

$$R_{a,b}(D) = \frac{1}{2N}\sum_{k=1}^{N}\max\left\{0\ ,\ \ln\left(\frac{\left[\sigma_k + \sqrt{\sigma_k^2 + [1-\frac{a}{b}]\alpha}\right]^2}{\alpha}\right)\right\}, \tag{4.61a}$$

$$D = \frac{1}{N}\sum_{k=1}^{N}\min\left\{b\sigma_k^2\ ,\ \frac{a\alpha\sigma_k/2}{\sigma_k + \sqrt{\sigma_k^2 + [1-\frac{a}{b}]\alpha}}\right\} \tag{4.61b}$$

*where*

$$\sigma_i^2 \triangleq \lambda_i(\boldsymbol{K}_{\mathbf{x}}), \quad i = 1, \ldots, N,$$

*are the eigenvalues of* $\boldsymbol{K}_{\mathbf{x}}$*. A reconstruction error* $\mathbf{z} = \mathbf{y} - \mathbf{x}$ *achieves* $R_{a,b}(D)$ *if and only if* $\mathbf{z}$ *is jointly Gaussian with* $\mathbf{x}$ *and if and only if the source-uncorrelated and source-parallel components of* $\mathbf{z}$ *are*

$$\boldsymbol{K}_{\mathbf{z}} - \boldsymbol{K}_{\mathbf{z},\mathbf{x}}\boldsymbol{K}_{\mathbf{x}}^{-1}\boldsymbol{K}_{\mathbf{z},\mathbf{x}}^{T} = \boldsymbol{K}_{\mathbf{u}}^{\star} \triangleq \boldsymbol{Q}\,\mathrm{diag}\left\{\max\left\{0\,,\ \frac{\alpha}{4}\left(1 - \frac{\alpha}{\left(\sigma_k + \sqrt{\sigma_k^2 + [1 - \frac{a}{b}]\alpha}\right)^2}\right)\right\}\right\}\boldsymbol{Q}^T \tag{4.62a}$$

$$\boldsymbol{K}_{\mathbf{z},\mathbf{x}}\boldsymbol{K}_{\mathbf{x}}^{-1}\boldsymbol{K}_{\mathbf{z},\mathbf{x}}^{T} = \boldsymbol{V}\boldsymbol{K}_{\mathbf{x}}\boldsymbol{V}^T = \boldsymbol{Q}\,\mathrm{diag}\left\{\min\left\{\sigma_k^2\,,\ \frac{(1/4)(a/b)^2\alpha^2}{\left(\sigma_k + \sqrt{\sigma_k^2 + [1 - \frac{a}{b}]\alpha}\right)^2}\right\}\right\}\boldsymbol{Q}^T \tag{4.62b}$$

*where* $\boldsymbol{Q}$ *is a unitary matrix having the eigenvectors of* $\boldsymbol{K}_{\mathbf{x}}$ *as its columns. The linear distortion matrix necessary to realize* $R_{a,b}(D)$ *is*

$$\boldsymbol{V}^{\star} \triangleq \boldsymbol{K}_{\mathbf{z},\mathbf{x}}\boldsymbol{K}_{\mathbf{x}}^{-1} = \boldsymbol{Q}\,\mathrm{diag}\left\{\min\left\{1\,,\ \frac{(1/2)(a/b)\alpha}{\sigma_k^2 + \sigma_k\sqrt{\sigma_k^2 + [1 - \frac{a}{b}]\alpha}}\right\}\right\}\boldsymbol{Q}^T. \tag{4.62c}$$

▲

*Proof.* The eigenvalue decomposition of $\boldsymbol{K}_{\mathbf{x}}$ is

$$\boldsymbol{K}_{\mathbf{x}} = \boldsymbol{Q}\,\mathrm{diag}\{\sigma_i^2\}\boldsymbol{Q}^T,$$

Since $\boldsymbol{Q}$ is invertible, the following holds

$$I(\mathbf{x}; \mathbf{x} + \mathbf{z}) = h(\mathbf{x}) - h(\mathbf{x}|\mathbf{x} + \mathbf{z}) = h(\boldsymbol{Q}^T\mathbf{x}) - h(\boldsymbol{Q}^T\mathbf{x}|\mathbf{x} + \mathbf{z}) = h(\boldsymbol{Q}^T\mathbf{x}) - h(\boldsymbol{Q}^T\mathbf{x}|\boldsymbol{Q}^T(\mathbf{x} + \mathbf{z})). \tag{4.63}$$

Define

$$\bar{\mathbf{x}} \triangleq \boldsymbol{Q}^T\mathbf{x}\,, \qquad \text{and} \qquad \bar{\mathbf{z}} \triangleq \boldsymbol{Q}^T\mathbf{z}. \tag{4.64}$$

Substitution of (4.64) into (4.63) yields

$$I(\mathbf{x}; \mathbf{x} + \mathbf{z}) = I(\bar{\mathbf{x}}; \bar{\mathbf{x}} + \bar{\mathbf{z}}). \tag{4.65}$$

Substituting (4.59) into (4.64) we obtain

$$\bar{\mathbf{z}} = \boldsymbol{Q}^T\mathbf{u} - \boldsymbol{Q}^T\boldsymbol{V}\mathbf{x} = \bar{\mathbf{u}} - \boldsymbol{Q}^T\boldsymbol{V}\boldsymbol{Q}\bar{\mathbf{x}} = \bar{\mathbf{u}} - \bar{\boldsymbol{V}}\bar{\mathbf{x}}, \tag{4.66a}$$

where

$$\bar{n} \triangleq \boldsymbol{Q}^T \mathbf{u}, \tag{4.66b}$$

$$\bar{\boldsymbol{V}} \triangleq \boldsymbol{Q}^T \boldsymbol{V} \boldsymbol{Q}, \text{ and} \tag{4.66c}$$

$$\mathrm{E}\left[\bar{\mathbf{x}}\bar{\mathbf{u}}^T\right] = 0.$$

Substitution of (4.66b) and (4.66c) into (4.60) yields

$$\begin{aligned} D_{a,b}(\mathbf{x}, \mathbf{x}+\mathbf{z}) &= \frac{1}{N} \operatorname{tr}\left\{a\boldsymbol{K}_{\mathbf{n}} + b\boldsymbol{V}\boldsymbol{K}_{\mathbf{x}}\boldsymbol{V}^T\right\} = \frac{1}{N} \operatorname{tr}\left\{a\boldsymbol{Q}^T\boldsymbol{K}_{\mathbf{n}}\boldsymbol{Q} + b\boldsymbol{Q}^T\boldsymbol{V}\boldsymbol{K}_{\mathbf{x}}(\boldsymbol{Q}^T\boldsymbol{V})^T\right\} \\ &= \frac{1}{N} \operatorname{tr}\left\{a\boldsymbol{Q}^T\boldsymbol{K}_{\mathbf{n}}\boldsymbol{Q} + b\boldsymbol{Q}^T\boldsymbol{V}\boldsymbol{Q}\boldsymbol{K}_{\bar{\mathbf{x}}}\boldsymbol{Q}^T\boldsymbol{V}^T\boldsymbol{Q}\right\} \\ &= \frac{1}{N} \operatorname{tr}\left\{a\boldsymbol{K}_{\bar{\mathbf{u}}} + b\bar{\boldsymbol{V}}\boldsymbol{K}_{\bar{\mathbf{x}}}\bar{\boldsymbol{V}}^T\right\}, \qquad (4.67) \\ &= D_{a,b}(\bar{\mathbf{x}}, \bar{\mathbf{x}}+\bar{\mathbf{z}}). \qquad (4.68) \end{aligned}$$

where (4.68) follows from (4.60) and (4.66). Therefore, in view of (4.65) and (4.68), the problem of finding the WCMSE-RDF for $\mathbf{x}$ is equivalent to finding the WCMSE-RDF for $\bar{\mathbf{x}}$. Conveniently, the latter is an i.i.d. Gaussian random vector with covariance matrix $\boldsymbol{K}_{\bar{\mathbf{x}}} = \operatorname{diag}\left\{\sigma_i^2\right\}$. Using the fact that $\boldsymbol{K}_{\mathbf{x}}$ is diagonal in (4.67) allows one to write $D_{a,b}(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ as

$$D_{a,b}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \frac{1}{N}\sum_{k=1}^{N} a\eta_k^2 + \frac{1}{N}\sum_{k=1}^{N} b\sigma_k^2 p_k^2 = \frac{1}{N}\sum_{k=1}^{N} d_k, \tag{4.69}$$

where

$$d_k \triangleq a\eta_k^2 + b\sigma_k^2 p_k^2 \qquad \eta_k^2 \triangleq [\boldsymbol{K}_{\bar{\mathbf{u}}}]_{k,k} \qquad p_k^2 \triangleq [\bar{\boldsymbol{V}}^T\bar{\boldsymbol{V}}]_{k,k}. \tag{4.70}$$

From these definitions and from (4.66), it is easy to see that $d_k$, $\eta_k^2$, and $p_k\sigma_k^2$ are, respectively, the WCMSE, the source-uncorrelated distortion, and the source-parallel distortion associated with the $k$-th scalar element of $\bar{\mathbf{x}}$. Denote the WCMSE rate-distortion function of the $k$-th element of $\bar{\mathbf{x}}$ by $r_k(d)$. From Theorem 4.2, we have

$$r_k(d) \triangleq \frac{1}{2}\ln\left(\max\left\{1\,,\ \frac{a}{d}\sigma_k^2 + 1 - \frac{a}{b}\right\}\right). \tag{4.71}$$

Recall from Lemma 4.3 that the rate-distortion functions $r_k(d)$ will be convex if and only if $a/b \leq 2$. From the fact that the elements of $\bar{\mathbf{x}}$ are i.i.d., and in view of (4.69), it is clear that $\bar{\mathbf{x}}$ is an $N$-fold product source and that $D_{a,b}(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a sum distortion metric. Thus, upon applying Theorem 4.5, we find that

$R_{a,b}(D)$ is given by

$$D = \frac{1}{N}\sum_{k=1}^{N} d_k,$$
$$R_{a,b}(D) = \frac{1}{N}\sum_{k=1}^{N} r_k(d_k), \tag{4.72}$$

where each scalar distortion $d_k$ is such that

$$\text{if } r_k'(b\sigma_k^2) > s, \qquad \text{then} \qquad r_k'(d_k) = s \tag{4.73}$$
$$\text{if } r_k'(b\sigma_k^2) \leq s, \qquad \text{then} \qquad d_k = b\sigma_k^2, \tag{4.74}$$

where $r_k'(t) = \frac{dr_k(t)}{dt}$, for some common slope $s < 0$. Differentiation of (4.71) yields

$$r_k'(d) = \begin{cases} -\dfrac{1/2}{d + [\frac{1}{a} - \frac{1}{b}]\frac{1}{\sigma_k^2} d^2} & , \; d \leq b\,\sigma_k^2 \\ 0 & , \; d > b\,\sigma_k^2 \end{cases} \qquad \forall k \in \{1, \ldots, N\} \tag{4.75}$$

Direct evaluation of the latter at $d = b\sigma_k^2$ yields that

$$r_k'(b\sigma_k^2) = -\frac{1/2}{b\sigma_k^2 + [\frac{b}{a} - 1]b\sigma_k^2} = -\frac{a}{2b^2\sigma_k^2},$$

which when replaced into (4.74) yields that

$$-\frac{a}{2b^2\sigma_k^2} \leq s \Longrightarrow d_k = b\sigma_k^2.$$

Else, if $-\frac{a}{2b^2\sigma_k^2} > s$, then the right-hand-side (RHS) of (4.73) must hold. In view of (4.75), the latter implies that $d_k$ satisfies

$$d_k + [\tfrac{1}{a} - \tfrac{1}{b}]\tfrac{1}{\sigma_k^2} d_k^2 = -1/(2s) \iff 0 = [\tfrac{1}{a} - \tfrac{1}{b}]d_k^2 + \sigma_k^2 d_k + \sigma_k^2/(2s). \tag{4.76}$$

If $[\frac{1}{a} - \frac{1}{b}] \neq 0$, then (4.76) holds iff

$$d_k = -\frac{\sigma_k^2 \pm \sigma_k\sqrt{\sigma_k^2 - 2[\frac{1}{a} - \frac{1}{b}]\frac{1}{s}}}{2[\frac{1}{a} - \frac{1}{b}]}. \tag{4.77}$$

In order to determine the correct sign preceding the square root in this equation, we recall from Lemma 4.3 that, if $a < 2b$, then $r_k(d)$ is a convex function, for every $k$. The convexity of $r_k(d)$, together with (4.73), implies that $\frac{d\,d_k}{ds} \geq 0$. To verify whether the latter holds, we differentiate (4.77) with respect to $s$. This yields

$$\frac{d\,d_k}{ds} = -\frac{\pm\sigma_k}{[\frac{1}{a} - \frac{1}{b}]} \cdot \frac{-2[\frac{1}{a} - \frac{1}{b}](-\frac{1}{s^2})}{\sqrt{\sigma_k^2 - 2[\frac{1}{a} - \frac{1}{b}]\frac{1}{s}}} = -\frac{\pm 2\sigma_k(\frac{1}{s^2})}{\sqrt{\sigma_k^2 - 2[\frac{1}{a} - \frac{1}{b}]\frac{1}{s}}}.$$

Thus, the only consistent option for the $\pm$ in (4.77) is the "minus" sign. Upon substituting $\pm$ by $-$, and then multiplying both the numerator and denominator by $\sigma_k + \sqrt{\sigma_k^2 - 2[\frac{1}{a} - \frac{1}{b}]\frac{1}{s}}$, (4.77) becomes

$$d_k = -\frac{\sigma_k \frac{1}{s}}{\sigma_k + \sqrt{\sigma_k^2 - [\frac{1}{a} - \frac{1}{b}]\frac{2}{s}}} = \frac{(1/2)a\alpha\sigma_k}{\sigma_k + \sqrt{\sigma_k^2 + [1 - \frac{a}{b}]\alpha}}, \tag{4.78}$$

where the change of variable

$$\alpha = -\frac{2}{as}$$

has been used. On the other hand, if $[\frac{1}{a} - \frac{1}{b}] = 0$, then (4.76) leads directly to $d_k = -1/(2s) = \frac{a\alpha}{4}$, which is the same result obtained from (4.78). Substitution of (4.78) into (4.71) and then into (4.72) yields (4.61a). Similarly, substitution of (4.78) into (4.69) yields (4.61b).

From (4.52) in Theorem 4.5, it also follows that $R_{a.b}(D)$ is achieved if and only if $\boldsymbol{K}_{\bar{\mathbf{z}}}$ and $\boldsymbol{K}_{\bar{\mathbf{z}},\bar{\mathbf{x}}}$ are diagonal matrices. Thus, from (4.70),

$$\boldsymbol{K}_{\bar{\mathbf{z}}} = \operatorname{diag}\left\{\eta_k^2 + p_k^2\sigma_k^2\right\}; \quad \boldsymbol{K}_{\bar{\mathbf{z}},\bar{\mathbf{x}}}\boldsymbol{K}_{\bar{\mathbf{x}}}^{-1} = -\operatorname{diag}\{p_k\}; \quad \boldsymbol{K}_{\bar{\mathbf{u}}} = \operatorname{diag}\left\{\eta_k^2\right\}. \tag{4.79}$$

From Theorem 4.2,

$$p_k = \frac{d_k}{b\sigma_k^2} = \frac{(1/2)(a/b)\alpha}{\sigma_k^2 + \sigma_k\sqrt{\sigma_k^2 + [1 - \frac{a}{b}]\alpha}},$$

and

$$\eta_k^2 = \frac{\alpha}{4}\left(1 - \frac{\alpha}{\left(\sigma_k + \sqrt{\sigma_k^2 + [1 - \frac{a}{b}]\alpha}\right)^2}\right).$$

Substitution of (4.80) into (4.79), together with the fact that

$$\boldsymbol{K}_{\mathbf{u}} = \boldsymbol{Q}\boldsymbol{K}_{\bar{\mathbf{u}}}\boldsymbol{Q}^T,$$
$$\boldsymbol{K}_{\mathbf{z},\mathbf{x}}\boldsymbol{K}_{\mathbf{x}}^{-1} = -\boldsymbol{V} = -\boldsymbol{Q}\bar{\boldsymbol{V}}\boldsymbol{Q}^T = \boldsymbol{Q}\boldsymbol{K}_{\bar{\mathbf{z}},\bar{\mathbf{x}}}\boldsymbol{K}_{\bar{\mathbf{x}}}^{-1}\boldsymbol{Q}^T,$$

yields (4.62). This completes the proof. □

**Remark 4.3.** *Strictly speaking, Theorem 4.6 characterizes $R_{a,b}(D)$ for i.i.d. vector processes only, since the proof of achievability to be provided in Section 4.8 requires the encoding of an infinite number of vectors.*

The characterization of the WCMSE rate-distortion function for vector sources given in Theorem 4.6 will be helpful in deriving the characterization of $R_{a,b}(D)$ for random stationary processes. This is done below.

## 4.5 WCMSE RDF For Gaussian Stationary Processes

For a w.s.s. random source $\{\mathrm{x}(k)\}$ reconstructed as the random process $\{\mathrm{y}(k)\}$, the reconstruction error is the process

$$\{\mathrm{z}(k)\} \triangleq \{\mathrm{y}(k)\} - \{\mathrm{x}(k)\}\,.$$

**Definition 4.3.** *The WCMSE-Rate Distortion Function for a w.s.s. source is defined as*

$$D_{a,b}(D) \triangleq \lim_{N\to\infty} \left( \min_{\{\mathrm{z}(k)\}: D_{a,b}(\mathbf{x}^{(N)},\, \mathbf{x}^{(N)}+\mathbf{z}^{(N)}) \le D} \bar{I}(\mathbf{x}^{(N)}; \mathbf{x}^{(N)} + \mathbf{z}^{(N)}) \right),$$

*where the random vectors*

$$\begin{aligned} \mathbf{x}^{(N)} &\triangleq \left[\mathrm{x}(-\tfrac{N-1}{2}), \mathrm{x}(-\tfrac{N-1}{2}+1), \ldots, \mathrm{x}(\tfrac{N-1}{2})\right]^T, \quad N \in \{2j+1\}_{j=0}^{\infty} \\ \mathbf{z}^{(N)} &\triangleq \left[\mathrm{z}(-\tfrac{N-1}{2}), \mathrm{z}(-\tfrac{N-1}{2}+1), \ldots, \mathrm{z}(\tfrac{N-1}{2})\right]^T, \quad N \in \{2j+1\}_{j=0}^{\infty}\,. \end{aligned}$$

▲

If $\{\mathrm{x}(k)\}$ and $\{\mathrm{z}(k)\}$ are jointly w.s.s., then the WCMSE takes the form

$$D_{a,b}(\{\mathrm{x}(k)\}, \{\mathrm{y}(k)\}) \triangleq a\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{\mathrm{u}}(\mathrm{e}^{j\omega})d\omega + b\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|V(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega})d\omega \tag{4.81}$$

where

$$V(\mathrm{e}^{j\omega}) \triangleq \frac{S_{\mathrm{z\,x}}(\mathrm{e}^{j\omega})}{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}, \quad \forall \omega \in [-\pi, \pi], \tag{4.82}$$

is the *linear distortion frequency response* and $S_{\mathbf{u}}(\mathrm{e}^{j\omega})$ is the PSD of the source-uncorrelated distortion, associated to the distortion $\{\mathrm{z}(k)\}$ and the source $\{\mathrm{x}(k)\}$.

The following theorem, which characterizes $R_{a,b}(D)$ for Gaussian stationary sources, also states that $R_{a,b}(D)$ is realized only when $\{\mathrm{z}(k)\}$ is a Gaussian stationary random process.

**Theorem 4.7** (WCMSE-RDF for Gaussian Stationary Processes)**.** *Let the source $\{\mathrm{x}(k)\}$ be a zero-mean Gaussian stationary random process with spectrum $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$ such that $S_{\mathrm{x}}(\mathrm{e}^{j\omega}) > 0$, a.e. on $[-\pi, \pi]$. Then*

*(i) For any $D > 0$,*

$$R_{a,b}(D) = \frac{1}{2\pi} \int\limits_{S_{\mathrm{x}} \geq \frac{a^2}{b^2}\alpha/4} \log\left( \frac{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [1-\frac{a}{b}]\alpha}}{\sqrt{\alpha}} \right) d\omega, \tag{4.83a}$$

*where $\alpha > 0$ is the only scalar parameter satisfying*

$$D = \frac{1}{2\pi} \int\limits_{S_{\mathrm{x}} \geq \frac{a^2}{b^2}\alpha/4} \frac{a\alpha\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}/2}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [1-\frac{a}{b}]\alpha}} d\omega + \frac{1}{2\pi} \int\limits_{S_{\mathrm{x}} < \frac{a^2}{b^2}\alpha/4} bS_{\mathrm{x}}(\mathrm{e}^{j\omega}) d\omega. \tag{4.83b}$$

*(ii) The mutual information rate $\bar{I}(\{\mathrm{x}\};\{\mathrm{x}(k)+\mathrm{z}(k)\}) = R_{a,b}(D)$ iff $\{\mathrm{z}(k)\}$ is stationary, and has the form*

$$\{\mathrm{z}(k)\} = \{\mathrm{u}(k)\} - p(k) * \{\mathrm{x}(k)\}, \tag{4.84}$$

*where $\{\mathrm{u}(k)\}$ is a Gaussian stationary random process independent of $\{\mathrm{x}(k)\}$ having PSD*

$$S^{\star}_{\mathrm{u}}(\mathrm{e}^{j\omega}) \triangleq \max\left\{ 0\ ,\ \frac{\alpha}{4}\left( 1 - \frac{\alpha}{\left( \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [1-\frac{a}{b}]\alpha} \right)^2} \right) \right\},\ \forall \omega \in [-\pi, \pi], \tag{4.85a}$$

*and where $\{p(k)\}_{k\in\mathbb{Z}}$ is a sequence of real numbers having discrete-time Fourier Transform*

$$V^{\star}(\mathrm{e}^{j\omega}) \triangleq \min\left\{ 1\ ,\ \frac{\frac{1}{2}(a/b)\alpha}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}\left( \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [1-\frac{a}{b}]\alpha} \right)} \right\},\ \forall \omega \in [-\pi, \pi]. \tag{4.85b}$$

▲

*Proof.* It is known that $\breve{\sigma}^{2^{(N)}} \geq \breve{\sigma}^{2^{(N+1)}}, \forall N \in \mathbb{N}$, where $\breve{\sigma}^{2^{(N)}}$ and $\breve{\sigma}^{2^{(N+1)}}$ are the smallest eigenvalues of the Toeplitz matrices $\boldsymbol{K}_{\mathbf{x}^{(N)}}$ and $\boldsymbol{K}_{\mathbf{x}^{(N+1)}}$, respectively (see e.g. [159, Theorem 4.3.8]). This result, together with Lemma 4.15, in the Appendix of this chapter, and the fact that $S_{\mathrm{x}}(\mathrm{e}^{j\omega}) > 0$, a.e. on $[-\pi, \pi]$, implies that $|\boldsymbol{K}_{\mathbf{x}^{(N)}}| > 0$, for all $N \in \mathbb{N}$. We can then apply Lemma 4.15 to (4.61), which yields

$$R_{a,b}(D) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \max\left\{ 0\ ,\ \log\left( \frac{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [1-\frac{a}{b}]\alpha}}{\sqrt{\alpha}} \right) \right\} d\omega, \tag{4.86a}$$

where $\alpha > 0$ is the only scalar parameter satisfying

$$D = \frac{1}{2\pi} \int_{-\pi}^{\pi} \min\left\{ bS_{\mathrm{x}}(\mathrm{e}^{j\omega})\ ,\ \frac{a\alpha\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}/2}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [1-\frac{a}{b}]\alpha}} \right\} d\omega. \tag{4.86b}$$

On the other hand,

$$\frac{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}+\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})+[1-\frac{a}{b}]\alpha}}{\sqrt{\alpha}}\geq 1 \iff S_{\mathrm{x}}(\mathrm{e}^{j\omega})\geq \frac{a^2}{b^2}\alpha/4 \tag{4.87a}$$

and

$$\frac{a\alpha\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}/2}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}+\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}+[1-\frac{a}{b}]\alpha}\leq bS_{\mathrm{x}}(\mathrm{e}^{j\omega}) \iff S_{\mathrm{x}}(\mathrm{e}^{j\omega})\geq \frac{a^2}{b^2}\alpha/4. \tag{4.87b}$$

By using (4.87), we have that (4.86) becomes (4.83).

In addition, it follows from Theorem 4.6 that a reconstruction error process $\{\mathrm{z}(k)\}$ realizes $R_{a,b}(D)$ if and only if and only if it is jointly Gaussian and jointly stationary with the source. This implies that $\{\mathrm{n}(k)\}$ in (4.84) is also a stationary Gaussian process. On the other hand, it is clear from (4.83) that $R_{a,b}(D)$ is determined only by the part of $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$ which stands above the threshold $\frac{a^2}{b^2}\alpha/4$. This implies that, in a realization of $R_{a,b}(D)$, the spectral components of the source outside the set of frequencies

$$\mathbb{B}\triangleq\left\{\omega\in[-\pi,\pi]: S_{\mathrm{x}}(\mathrm{e}^{j\omega})>\frac{a^2}{b^2}\alpha/4\right\}$$

must be suppressed, while spectral components of the source within the set of frequencies $\mathbb{B}$ suffer distortion, but are not absent in the reconstruction. It is also evident that, in a realization of $R_{a,b}(D)$, the reconstructed process $\{\mathrm{y}(k)\}\triangleq\{\mathrm{x}(k)\}+\{\mathrm{z}(k)\}$ must have a PSD which is zero $\forall\omega\notin\mathbb{B}$. From this, it follows that if the statistics of the reconstruction error $\{\mathrm{z}(k)\}$ realize $R_{a,b}(D)$, then the mutual information rate between source and reconstruction can be written as

$$\bar{I}(\{\mathrm{x}(k)\};\{\mathrm{y}(k)\})=\frac{1}{4\pi}\int\limits_{\omega\in\mathbb{B}}\log\left(\frac{\left|1-V(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega})}{S_{\mathrm{u}}(\mathrm{e}^{j\omega})}+1\right)d\omega, \tag{4.88}$$

see[5] Fact 2.4 in Section 2.3. In (4.88),

$$\begin{aligned} V(\mathrm{e}^{j\omega})&\triangleq\frac{S_{\mathrm{z,x}}(\mathrm{e}^{j\omega})}{S_{\mathrm{x}}(\mathrm{e}^{j\omega})},\\ S_{\mathrm{u}}(\mathrm{e}^{j\omega})&\triangleq S_{\mathrm{z}}(\mathrm{e}^{j\omega})-\frac{S_{\mathrm{z,x}}(\mathrm{e}^{j\omega})^2}{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}\end{aligned} \tag{4.89}$$

are, respectively, the linear-distortion frequency response associated to $\{\mathrm{z}(k)\}$ and $\{\mathrm{x}(k)\}$, and the PSD of the source-uncorrelated distortion component of $\{\mathrm{z}(k)\}$. The fact that no spectral components of the source associated with frequencies in $\mathbb{B}$ are suppressed implies that

$$V(\mathrm{e}^{j\omega})\neq 1,\quad \forall\omega\in\mathbb{B},$$

[5]The expression for $\bar{I}(\{\mathrm{x}(k)\};\{\mathrm{y}(k)\})$ in (4.88) can be obtained by grouping the spectral components of $S_{\mathrm{u}}(\mathrm{e}^{j\omega})$ and $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$ into a single continuous band and then critically decimating the result to obtain PSDs which are non-zero over $[-\pi,\pi]$. In doing this, the mutual information rate between $\{\mathrm{x}(k)\}$ and $\{\mathrm{z}(k)\}$ is preserved, and (2.40) leads to (4.88).

and thus, from (4.89), we obtain that $S_{\mathrm{u}}(\mathrm{e}^{j\omega}) > 0$, $\forall \omega \in \mathbb{B}$. In turn, the WCMSE is given by

$$D_{a,b}(\{\mathrm{x}(k)\}, \{\mathrm{y}(k)\}) = \frac{1}{2\pi} \int\limits_{\omega \in \mathbb{B}} a S_{\mathrm{u}}(\mathrm{e}^{j\omega}) + b \left|V(\mathrm{e}^{j\omega})\right|^2 d\omega + \frac{1}{2\pi} \int\limits_{\omega \notin \mathbb{B}} b S_{\mathrm{x}}(\mathrm{e}^{j\omega}) d\omega. \tag{4.90}$$

It is clear from (4.88) and (4.90) that an optimal $V(\mathrm{e}^{j\omega})$, i.e., one that minimizes $\bar{I}(\{\mathrm{x}(k)\}; \{\mathrm{y}(k)\})$ for a given $D_{a,b}(\{\mathrm{x}(k)\}, \{\mathrm{y}(k)\})$, must be positive and real for all $\omega$. Let

$$S_{\mathrm{d}}(\mathrm{e}^{j\omega}) \triangleq a S_{\mathrm{u}}(\mathrm{e}^{j\omega}) + b \left|V(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega}), \tag{4.91}$$

and define

$$D^{\mathbb{B}}_{a,b} \triangleq \frac{1}{2\pi} \int\limits_{\omega \in \mathbb{B}} \frac{a\alpha \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}/2}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [1 - \frac{a}{b}]\alpha}} d\omega = \frac{1}{2\pi} \int\limits_{\omega \in \mathbb{B}} S_{\mathrm{d}}(\mathrm{e}^{j\omega}) d\omega \tag{4.92}$$

For any given $S_{\mathrm{d}}(\mathrm{e}^{j\omega})$, it follows from (4.88), (4.92) and Theorem 4.2, and from the fact that $V(\mathrm{e}^{j\omega}) < 1$, $\forall \omega \in \mathbb{B}$, that the optimal $S_{\mathrm{u}}(\mathrm{e}^{j\omega})$ and $V(\mathrm{e}^{j\omega})$ satisfy

$$S_{\mathrm{u}}(\mathrm{e}^{j\omega}) = \frac{S_{\mathrm{d}}(\mathrm{e}^{j\omega})}{a} \left(1 - \frac{S_{\mathrm{d}}(\mathrm{e}^{j\omega})}{b S_{\mathrm{x}}(\mathrm{e}^{j\omega})}\right), \qquad \forall \omega \in \mathbb{B} \tag{4.93}$$

$$V(\mathrm{e}^{j\omega}) = \frac{S_{\mathrm{d}}(\mathrm{e}^{j\omega})}{b S_{\mathrm{x}}(\mathrm{e}^{j\omega})}, \qquad \forall \omega \in \mathbb{B}. \tag{4.94}$$

Substitution of (4.93) and (4.94) into (4.88) yields

$$\begin{aligned} \bar{I}(\{\mathrm{x}(k)\}; \{\mathrm{y}(k)\}) &= \frac{1}{4\pi} \int\limits_{\omega \in \mathbb{B}} \log \left( \frac{\left(1 - \frac{S_{\mathrm{d}}(\mathrm{e}^{j\omega})}{b S_{\mathrm{x}}(\mathrm{e}^{j\omega})}\right)^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega})}{\frac{S_{\mathrm{d}}(\mathrm{e}^{j\omega})}{a}\left(1 - \frac{S_{\mathrm{d}}(\mathrm{e}^{j\omega})}{b S_{\mathrm{x}}(\mathrm{e}^{j\omega})}\right)} + 1 \right) d\omega \\ &= \frac{1}{4\pi} \int\limits_{\omega \in \mathbb{B}} \log \left( \frac{a S_{\mathrm{x}}(\mathrm{e}^{j\omega})}{S_{\mathrm{d}}(\mathrm{e}^{j\omega})} + 1 - \frac{a}{b} \right) d\omega. \end{aligned} \tag{4.95}$$

Let us now find the optimal $S_{\mathrm{d}}(\mathrm{e}^{j\omega})$. The latter needs to be such that it minimizes (4.95) subject to the constraint (4.92). In order to find such minimizer, we define, in relation to (4.95) and (4.92), the Lagrangian

$$\mathscr{L} \triangleq \log \left( \frac{a S_{\mathrm{x}}(\mathrm{e}^{j\omega})}{S_{\mathrm{d}}(\mathrm{e}^{j\omega})} + 1 - \frac{a}{b} \right) + \lambda S_{\mathrm{d}}(\mathrm{e}^{j\omega}), \tag{4.96}$$

and notice that the optimal $S_{\mathrm{d}}(\mathrm{e}^{j\omega})$ is such that $\frac{\partial \mathscr{L}}{\partial S_{\mathrm{d}}}$ is zero for some multiplier $\lambda > 0$ and for all $\omega$.

Differentiating (4.96),

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial S_{\mathrm{d}}} &= -\frac{\frac{aS_{\mathrm{x}}(\mathrm{e}^{j\omega})}{S_{\mathrm{d}}(\mathrm{e}^{j\omega})^2}}{\frac{aS_{\mathrm{x}}(\mathrm{e}^{j\omega})}{S_{\mathrm{d}}(\mathrm{e}^{j\omega})}+1-\frac{a}{b}} + \lambda = 0 \\
&\iff -\frac{aS_{\mathrm{x}}(\mathrm{e}^{j\omega})}{aS_{\mathrm{x}}(\mathrm{e}^{j\omega})S_{\mathrm{d}}(\mathrm{e}^{j\omega}) + [1-\frac{a}{b}]S_{\mathrm{d}}(\mathrm{e}^{j\omega})^2} + \lambda = 0 \\
&\iff [\tfrac{1}{a}-\tfrac{1}{b}]S_{\mathrm{d}}(\mathrm{e}^{j\omega})^2 + S_{\mathrm{x}}(\mathrm{e}^{j\omega})S_{\mathrm{d}}(\mathrm{e}^{j\omega}) - \tfrac{1}{\lambda}S_{\mathrm{x}}(\mathrm{e}^{j\omega}) = 0 \\
\iff S_{\mathrm{d}}(\mathrm{e}^{j\omega}) &= \frac{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}}{2[\frac{1}{a}-\frac{1}{b}]}\left(-\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} \pm \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [\tfrac{1}{a}-\tfrac{1}{b}]\tfrac{4}{\lambda}}\right) \\
&= \frac{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}}{2[\frac{1}{a}-\frac{1}{b}]}\left(\frac{-[\frac{1}{a}-\frac{1}{b}]\frac{4}{\lambda}}{-\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} \mp \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [\frac{1}{a}-\frac{1}{b}]\frac{4}{\lambda}}}\right) \\
&= \frac{\frac{2}{\lambda}\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} \pm \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [\frac{1}{a}-\frac{1}{b}]\frac{4}{\lambda}}} \qquad (4.97) \\
&= \frac{\frac{2}{\lambda}\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [\frac{1}{a}-\frac{1}{b}]\frac{4}{\lambda}}}, \qquad (4.98)
\end{aligned}$$

where (4.98) follows from the fact that choosing "minus" for the $\pm$ sign in (4.97) yields $S_{\mathrm{d}} \geq bS_{\mathrm{x}}(\mathrm{e}^{j\omega})$, $\forall\omega \in [-\pi,\pi]$, which would contradict the optimality of $S_{\mathrm{d}}(\mathrm{e}^{j\omega})$. Equating (4.98) with (4.92), it becomes clear that $\frac{4}{a\lambda} = \alpha$. Thus, the optimal $S_{\mathbf{d}}(\mathrm{e}^{j\omega})$ is given by

$$S_{\mathrm{d}}(\mathrm{e}^{j\omega}) = \frac{\frac{1}{2}a\alpha\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [\frac{1}{a}-\frac{1}{b}]\frac{4}{\lambda}}} \qquad (4.99)$$

Substitution of (4.99) into (4.94) shows that the linear distortion frequency response that realizes $R_{a,b}(D)$ is given by

$$\begin{aligned}
V(\mathrm{e}^{j\omega}) &= \frac{\frac{\frac{1}{2}a\alpha\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}+\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})+[\frac{1}{a}-\frac{1}{b}]\frac{4}{\lambda}}}}{bS_{\mathrm{x}}(\mathrm{e}^{j\omega})}, \qquad \forall\omega \in \mathbb{B}. \\
&= \frac{\frac{1}{2}(a/b)\alpha}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}\left(\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [\frac{1}{a}-\frac{1}{b}]\frac{4}{\lambda}}\right)}
\end{aligned}$$

Similarly, the source uncorrelated noise present in $\{\mathrm{z}(k)\}$ that realizes $R_{a,b}(D)$ is obtained by substituting (4.99) into (4.93), which gives

$$S_{\mathrm{u}}(\mathrm{e}^{j\omega}) = \frac{\alpha}{4}\left(1 - \frac{\alpha}{\left(\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [1-\frac{a}{b}]\alpha}\right)^2}\right), \quad \forall\omega \in \mathbb{B}.$$

These equations, together with the fact that $V(\mathrm{e}^{j\omega}) = 1, \forall\omega \notin \mathbb{B}$, lead directly to (4.85). This completes the proof. □

**Remark 4.4.** *By comparing* (4.83) *with* (3.118) *and* (3.121)*, we see that, for any WCMSE value $D$, the SNR of $\mathcal{Q}$ in an optimal feedback scalar quantizer satisfies*

$$\frac{1}{2}\log(\gamma+1) = R_{a,b}(D). \tag{4.100}$$

*In view of* (2.60) *(see page 42), this implies that, for Gaussian sources and using entropy coded scalar quantization with dither and optimal filters, the operational rate exceeds $R_{a,b}(D)$ by less than $0.254$ bits/sample. This is a generalization of the result obtained in [15], where the MSE distortion criterion is used, to the WCMSE distortion criterion.* ▲

### 4.5.1 Distortion Spectra

It is well known that in a realization of Shannon's rate-distortion function for a Gaussian stationary process source $\{\mathrm{x}(k)\}$ and MSE as the distortion metric, the PSD of the reconstruction error, $S_{\mathrm{z}}(\mathrm{e}^{j\omega})$, is constant over the frequency bands in which $S_{\mathrm{z}}(\mathrm{e}^{j\omega}) \leq S_{\mathrm{x}}(\mathrm{e}^{j\omega})$ (see (1.1) and Fig. 1.1 in Section 1.1). As will be discussed next, this is not the case for $R_{a,b}(D)$, in general.

From (4.84), the PSD of the reconstruction error process $\{\mathrm{z}(k)\}$ is

$$S_{\mathrm{z}}(\mathrm{e}^{j\omega}) = S_{\mathrm{u}}(\mathrm{e}^{j\omega}) + \left|V(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega}) \tag{4.101}$$

Substituting (4.85) in to this equation, and restricting hereafter, for simplicity, to the cases in which $S_{\mathrm{x}}(\mathrm{e}^{j\omega}) \geq \frac{1}{4}(a/b)^2\alpha,$ a.e. on $[-\pi,\pi]$, we obtain

$$S_{\mathrm{z}}(\mathrm{e}^{j\omega}) = \frac{\alpha}{4} + \frac{\left[(\frac{a}{b})^2 - 1\right]\alpha^2}{\left(\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + [1-\frac{a}{b}]\alpha}\right)^2} \tag{4.102}$$

From this equation it is clear that, unless $a = b$, the PSD of $S_{\mathrm{z}}(\mathrm{e}^{j\omega})$ in a realization of $R_{a,b}(D)$ is not constant. This is not surprising, in view of the fact that, in the WCMSE, source-uncorrelated distortion (which gives rise to $S_{\mathrm{u}}(\mathrm{e}^{j\omega})$), has a different importance than source-parallel distortion (given by $\left|V(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega})$). On the other hand, this reasoning would suggest that the weighted sum

$$S_{\mathrm{d}}(\mathrm{e}^{j\omega}) = aS_{\mathrm{u}}(\mathrm{e}^{j\omega}) + \left|V(\mathrm{e}^{j\omega})\right| S_{\mathrm{x}}(\mathrm{e}^{j\omega}) \tag{4.103}$$

(already defined in (4.91)), is constant. This would seem reasonable, since, in $R_{a,b}(D)$, the distortion metric regards source-uncorrelated distortion power as being $a/b$ times more "expensive" than source-parallel error power. In terms of the geometric interpretation given in Section 4.3.1, scaling $D^{\perp}$ by $a$ and $D^{\parallel}$ by $b$ would turn all the ellipses in Fig. 4.1-(b) into circles, which is reminiscent of what one obtains when MSE is the distortion metric. However, and perhaps surprisingly, it can be seen from (4.99), that the weighted spectrum in (4.103) is not constant either.

Nevertheless, there exists a weighted combination of $S_{\mathrm{u}}(\mathrm{e}^{j\omega})$ and $\left|V(\mathrm{e}^{j\omega})\right| S_{\mathrm{x}}(\mathrm{e}^{j\omega})$ that is constant over $[-\pi,\pi]$ (again, assuming $S_{\mathrm{x}}(\mathrm{e}^{j\omega}) \geq \frac{1}{4}(a/b)^2\alpha,\ \forall\omega\in[-\pi,\pi]$). It is the following:

$$S_{\mathrm{u}}(\mathrm{e}^{j\omega}) + \left(\tfrac{b}{a}\right)^2 \left|V(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega}) = \frac{\alpha}{4}, \tag{4.104}$$

which can be readily verified by substituting (4.85) into the left-hand side of (4.104). Interestingly, in the weighted sum of spectral densities in (4.104), the ratio of weights of source-uncorrelated/parallel powers is the *square* of the ratio $a/b$.

### 4.5.2 Special Cases

**The Reverse Water-Filling Equations.** As already noted in Section 1.3.1, the WCMSE becomes the standard MSE when $a = b = 1$, that is, for $N$-dimensional random vectors $\mathbf{x}$ and $\mathbf{y}$, $D_{1,1}(\mathbf{x}.\mathbf{y}) = \frac{1}{N}\mathrm{E}\left[\|\mathbf{y}-\mathbf{x}\|^2\right] = \mathrm{MSE}$. As a consequence, $R_{1,1}(D)$ corresponds to Shannon's rate-distortion function when MSE is the distortion metric. This can be easily verified from Theorems 4.2, 4.6 and 4.7. For example, for Gaussian stationary random sources, (4.83) becomes the well known reverse water-filling equations, described in Section 1.1, with the "water-level" $\theta$ being $\frac{\alpha}{4}$.

**The Quadratic Gaussian RDF for Source-Uncorrelated Distortions.** On the other hand, by letting $b\to\infty$, all realizations of $R_{a,b}(D)$ are such that the reconstruction error is Gaussian and independent of the source (see Theorems 4.2, 4.6 and 4.7). The rate-distortion function corresponding to the case $a=1$ and $b\to\infty$ has been recently introduced and characterized by the author as the *quadratic Gaussian rate-distortion function for source uncorrelated distortions*, denoted by $R^{\perp}(D)$ [127]. We can define $R^{\perp}(D)$ in terms of $R_{a,b}(D)$ as

$$R^{\perp}(D) \triangleq R_{1,\infty}(D),$$

where $R_{1,\infty}(D) \triangleq \lim_{b\to\infty} R_{1,b}(D)$. Alternatively, $R^{\perp}(D)$ can be defined as follows

**Definition 4.4.** *The uncorrelated quadratic rate-distortion function $R^{\perp}(D)$ for a random process source $\{\mathrm{x}(k)\}$ is defined as*

$$R^{\perp}(D) = \min_{\substack{\{\mathrm{z}(k)\}\,:\,E[\mathrm{x}(j)\,\mathrm{z}(k)]=0,\,\forall j,k\in\mathbb{Z},\\ \lim_{N\to\infty}\frac{1}{2N+1}\sum_{k=-N}^{N}E\left[\mathrm{z}(k)^2\right]\leq D,}} \bar{I}(\{\mathrm{x}(k)\}\,;\{\mathrm{x}(k)+\mathrm{z}(k)\}), \tag{4.105}$$

*where $\{\mathrm{z}(k)\}$ is a random process.* ▲

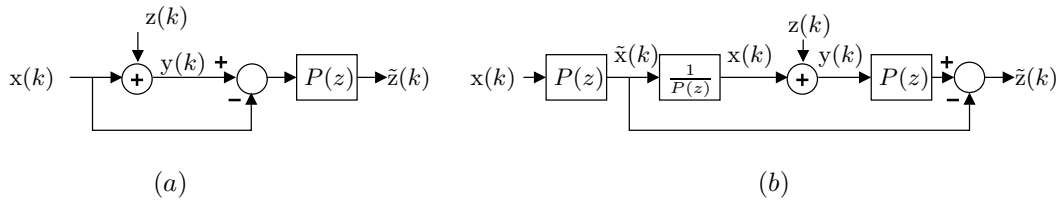



**Figure 4.4:** (a): Frequency weighting of the error $\mathrm{z}(k)$; (b): Equivalent scheme.

From (4.83), and taking $a = 1$, $b = \infty$, we obtain

$$R^{\perp}(D) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left( \frac{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + \alpha}}{\sqrt{\alpha}} \right) d\omega, \tag{4.106}$$

where $\alpha > 0$ is the only scalar parameter satisfying

$$D = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\alpha \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}}{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + \alpha}} d\omega. \tag{4.107}$$

Also, $R^{\perp}(D)$ is achieved if and only if the error $\{\mathrm{z}(k)\}$ is a Gaussian stationary process, independent of the source, and having PSD

$$S_{\mathrm{z}}(\mathrm{e}^{j\omega}) = \frac{1}{2} \left( \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + \alpha} - \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} \right) \sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}, \quad \text{a.e. on } [-\pi, \pi]. \tag{4.108}$$

$R^{\perp}(D)$ can be easily extended to consider frequency weighting of the source-uncorrelated error $\{\mathrm{z}(k)\}$. For this purpose, consider the setting depicted in Fig. 4.4-(a). In this scheme, the error frequency weighting filter $P(z)$ is bi-proper, stable, and stably invertible. The sequence $\{\tilde{\mathrm{z}}(k)\}$ is the frequency weighted reconstruction error. Associated to $P(z)$, we define define the frequency weighted distortion metric

$$J(P(z), \{\mathrm{z}(k)\}) \triangleq \lim_{\ell \to \infty} \frac{1}{\ell} \sum_{k=1}^{\ell} \mathrm{E}\left[\tilde{\mathrm{z}}(k)^2\right], \quad \text{where} \tag{4.109}$$

$$\tilde{\mathrm{z}}(k) \triangleq P(z)\, \mathrm{z}(k), \quad \forall k \in \mathbb{Z}. \tag{4.110}$$

An equivalent scheme is shown in Fig. 4.4-(b). Since $P(z)$ is invertible, we have that

$$\bar{I}(\{\tilde{\mathrm{x}}(k)\}; \{\tilde{\mathrm{x}}(k) + \mathrm{u}(k)\}) = \bar{I}(\{\mathrm{x}(k)\}; \{\mathrm{x}(k) + \mathrm{z}(k)\}). \tag{4.111}$$

Thus, under a constraint on the maximum average power of $\{\tilde{\mathrm{z}}(k)\}$, minimizing $\bar{I}(\{\tilde{\mathrm{x}}(k)\}; \{\tilde{\mathrm{x}}(k) + \tilde{\mathrm{z}}(k)\})$ over all processes $\{\tilde{\mathrm{z}}(k)\}$ is equivalent to minimizing $\bar{I}(\{\mathrm{x}(k)\}; \{\mathrm{x}(k) + \mathrm{z}(k)\})$ over all processes $\{\mathrm{z}(k)\}$. Taking $\{\tilde{\mathrm{x}}(k)\}$ as the source and $\{\tilde{\mathrm{z}}(k)\}$ as the noise in (4.106) and (4.107), and noting that

$S_{\tilde{\mathrm{x}}}(\mathrm{e}^{j\omega}) = \left|P(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega})$, it follows that the frequency weighted version of $R^{\perp}(D)$ is

$$R^{\perp}(D) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\frac{\left|P(\mathrm{e}^{j\omega})\right|\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{\left|P(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + \alpha}}{\sqrt{\alpha}}\right) d\omega, \tag{4.112}$$

where $\alpha > 0$ is the only scalar parameter satisfying

$$D = \frac{1}{4\pi}\int_{-\pi}^{\pi} \frac{\alpha\left|P(\mathrm{e}^{j\omega})\right|\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}}{\left|P(\mathrm{e}^{j\omega})\right|\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})} + \sqrt{\left|P(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + \alpha}} d\omega. \tag{4.113}$$

Also, since $S_{\tilde{\mathrm{z}}}(\mathrm{e}^{j\omega}) = \left|P(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{z}}(\mathrm{e}^{j\omega})$, we have from (4.108) that $R^{\perp}(D)$ is achieved if and only if the error $\{\mathrm{z}(k)\}$ is a Gaussian stationary process, independent of the source, and having PSD

$$S_{\mathrm{z}}(\mathrm{e}^{j\omega}) = \frac{1}{2}\left(\sqrt{\left|P(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + \alpha} - \left|P(\mathrm{e}^{j\omega})\right|\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}\right)\frac{\sqrt{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}}{|P(\mathrm{e}^{j\omega})|}, \quad \text{a.e. on } [-\pi,\pi]. \tag{4.114}$$

Notice that the expression for the frequency weighted MSE given by (4.113) is identical to that in (3.143), obtained for the optimal perfect reconstruction feedback quantizer under the Linear Model. Notice also that here, again, $\frac{1}{2}\log(\gamma+1)$ in (3.142) plays the same role as the RDF, this time $R^{\perp}(D)$ in (4.112). (This correspondence will be studied in detail and extended in Chapter 5.) From this and from (2.60) (page 42), it follows that for Gaussian sources, an FQ using an entropy coded scalar quantizer with subtractive and the optimal filters characterized by (3.140), attains an operational bit-rate that exceeds $R^{\perp}(D)$ by less than $0.254$ bits/sample.

One of the attractive aspects of $R^{\perp}(D)$ is that it is the asymptotically achievable lower bound to the bit-rate of *any* ED pair satisfying the following two conditions: a) quantization errors are uncorrelated to the source; and b) the transfer function from source to reconstruction is unity (perfect-reconstruction property) [127, 160]. In particular, whenever a PR subband coder is analyzed assuming quantization errors uncorrelated to the source (the Linear Model), the difference between the bit-rate of the subband coder and $R^{\perp}(D)$, for a given distortion $D$, can be used as a measure of rate-distortion efficiency. This performance assessment criterion can be seen as an alternative to the *coding gain* criterion. The latter has been widely used in subband coding literature, e.g., see [55, 59, 86]. Unlike the coding gain, the performance gap of a PR source coder, satisfying condition (a), with respect to $R^{\perp}(D)$, is an absolute rate-distortion efficiency measure. This measure, instead of telling one how much better than PCM a given source coder is, tells one how far the latter is from the best conceivable source coder.

Another interesting feature of $R^{\perp}(D)$ is that, as shown by the author in [160], $R^{\perp}(D)$ can be realized

using only causal filters. Moreover, by using entropy-coded subtractively dithered scalar quantization, it is possible to obtain a bit-rate that exceeds $R^{\perp}(D)$ by not more than $0.254$ bits/sample, with zero delay from source to reconstruction [160]. This property makes it possible to extend $R^{\perp}(D)$ to situations wherein linear, time invariant feedback, exists between reconstruction and source. This will be the subject of Section 4.9.

It is easy to verify that $R^{\perp}(D)$ is the only rate-distortion function within the $R_{a,b}(D)$ family that can be realized causally. To see this, notice that the optimal linear distortion frequency response $V(\mathrm{e}^{j\omega})$ in (4.85b) is real valued and symmetric. Hence, the signal transfer function necessary to realize $R_{a,b}(D)$, which has frequency response $1 - V(\mathrm{e}^{j\omega})$, can only be implemented using non-causal filters. The only exception arises when $b \to \infty$. In this case, the optimal $V(\mathrm{e}^{j\omega}) \equiv 0$ (see (4.85b)), and the optimal signal transfer function is unity, which can always be realized with causal filters.

## 4.6 WCMSE-RDF For Vector Processes

Here we extend the results of sections 4.3 and 4.4 to vector processes. For a Gaussian stationary vector process source $\{\mathbf{x}(k)\}$, where $\mathbf{x}(k) \in \mathbb{R}^N$, $\forall k \in \mathbb{Z}$, reconstructed as the process $\{\mathbf{y}(k)\}$, the reconstruction error is the process

$$\{\mathbf{z}(k)\} \triangleq \{\mathbf{y}(k) - \mathbf{x}(k)\}.$$

If $\{\mathbf{x}(k)\}$ and $\{\mathbf{z}(k)\}$ are jointly w.s.s., then the source-uncorrelated distortion is given by

$$\{\mathbf{u}(k)\} \triangleq \{\mathbf{z}\} + \boldsymbol{V}(z)\{\mathbf{x}(k)\}, \tag{4.115}$$

where the *linear distortion transfer function matrix* $\boldsymbol{V}(z)$ is defined as

$$\boldsymbol{V}(z) \triangleq -\boldsymbol{K}_{\mathbf{z},\mathbf{x}}(z)\boldsymbol{K}_{\mathbf{x}}(z)^{-1}. \tag{4.116}$$

From (4.115), the source-uncorrelated and source-parallel distortion terms that comprise the covariance matrix of $\{\mathbf{z}(k)\}$ are readily found to be, respectively,

$$\begin{aligned}\boldsymbol{D}^{\perp} &\triangleq \boldsymbol{K}_{\mathbf{u}},\\ \boldsymbol{D}^{\parallel} &\triangleq \boldsymbol{V}(z)\boldsymbol{K}_{\mathbf{x}}(k)\boldsymbol{V}(z)^{H} = \boldsymbol{K}_{\mathbf{z},\mathbf{x}}(z)\boldsymbol{K}_{\mathbf{x}}(z)\boldsymbol{K}_{\mathbf{z},\mathbf{x}}(z)^{H}.\end{aligned}$$

Although for vector process sources the reconstruction error that realizes $R_{a,b}(D)$ turns out to be jointly stationary with the source (as will be shown in Theorem 4.8), we cannot "a priori" assume stationarity when defining $R_{a,b}(D)$. For this purpose, we introduce the following notation for a concatenation

of the vectors $\{\mathbf{x}(k)\}_{k=-\ell}^{\ell}$ of a process $\{\mathbf{x}(k)\}$ into a single vector of length $(2\ell+1)N$:

$$\overline{\mathbf{x}}(\ell) \triangleq \left[\mathbf{x}(-\ell)^T \; \mathbf{x}(-\ell+1)^T \; \cdots \; \mathbf{x}(\ell)^T\right]^T. \tag{4.117}$$

Using this notation, the WCMSE for vector sources is defined as the following limit:

$$D_{a,b}(\{\mathbf{x}(k)\},\{\mathbf{y}(k)\} \triangleq \lim_{\ell\to\infty} D_{a,b}(\overline{\mathbf{x}}(k),\overline{\mathbf{y}}(k)), \tag{4.118}$$

where $D_{a,b}(\overline{\mathbf{x}}(k),\overline{\mathbf{y}}(k))$ corresponds to the WCMSE for vectors defined in (4.60).

We can now define the WCMSE-Rate-Distortion function for a vector process source $\{\mathbf{x}(k)\}$, as follows.

**Definition 4.5.** *For a vector random process source $\{\mathbf{x}(k)\}$, the WCMSE-RDF is defined as*

$$R_{a,b}(D) \triangleq \lim_{\ell\to\infty} \min_{\overline{\mathbf{z}}(\ell):D_{a,b}(\overline{\mathbf{x}}(\ell),\overline{\mathbf{x}}(\ell)+\overline{\mathbf{z}}(\ell))\le D} \bar{I}\left(\overline{\mathbf{x}}(\ell)\,;\,\overline{\mathbf{x}}(\ell)+\overline{\mathbf{z}}(\ell)\right). \tag{4.119}$$

▲

**Theorem 4.8.** *Let $\{\mathbf{x}(k)\}$ be a Gaussian stationary $N$-dimensional vector process with zero mean and covariance matrix $\boldsymbol{K}_{\mathbf{x}}(z)$. Let the eigenvalue decomposition of $\boldsymbol{K}_{\mathbf{x}}(z)$ be*

$$\boldsymbol{K}_{\mathbf{x}}(z) = \boldsymbol{Q}(z)\,\mathrm{diag}\left\{\lambda_i(z)\right\}\boldsymbol{Q}(z)^H, \tag{4.120}$$

*where $\{\lambda_i(z)\}_{i=1}^N$ are the eigen-functions of $\boldsymbol{K}_{\mathbf{x}}(z)$ and $\boldsymbol{Q}(z)$ is a unitary transfer function. Then*

$$R_{a,b}(D) = \sum_{i=1}^{N}\frac{1}{2\pi}\int_{\lambda_i(\omega)\ge\frac{a^2}{b^2}\alpha/4} \ln\left(\frac{\sqrt{\lambda_i(\omega)} + \sqrt{\lambda_i(\omega)+[1-\frac{a}{b}]\alpha}}{\sqrt{\alpha}}\right)d\omega, \tag{4.121a}$$

*where $\alpha > 0$ is the unique scalar satisfying*

$$D = \frac{1}{2\pi}\sum_{i=1}^{N}\left(\int_{\lambda_i(\omega)\ge\frac{a^2}{b^2}\alpha/4} \frac{a\alpha\sqrt{\lambda_i(\omega)}/2}{\sqrt{\lambda_i(\omega)} + \sqrt{\lambda_i(\omega)+[1-\frac{a}{b}]\alpha}}d\omega + \int_{\lambda_i(\omega)<\frac{a^2}{b^2}\alpha/4} b\lambda_i(\omega)d\omega\right). \tag{4.121b}$$

*In addition, $R_{a,b}(D)$ is achieved if and only if the source-uncorrelated distortion has covariance matrix*

$$K_{\mathbf{u}}^{\star}(\mathrm{e}^{j\omega}) = \boldsymbol{Q}(\mathrm{e}^{j\omega})\,\mathrm{diag}\left\{S_{\mathrm{u}_i}^{\star}(\mathrm{e}^{j\omega})\right\}\boldsymbol{Q}(\mathrm{e}^{j\omega})^H, \tag{4.121c}$$

*and the frequency response of the linear distortion transfer matrix (see (4.116)) is*

$$\boldsymbol{V}^{\star}(\mathrm{e}^{j\omega}) = \boldsymbol{Q}(\mathrm{e}^{j\omega})\,\mathrm{diag}\left\{V_i^{\star}(\mathrm{e}^{j\omega})\right\}\boldsymbol{Q}(\mathrm{e}^{j\omega})^H, \tag{4.121d}$$

*where*

$$S^{\star}_{\mathrm{u}_i}(\mathrm{e}^{j\omega}) \triangleq \max\left\{0\ ,\ \frac{\alpha}{4}\left(1-\frac{\alpha}{\left(\sqrt{\lambda_i(\mathrm{e}^{j\omega})}+\sqrt{\lambda_i(\mathrm{e}^{j\omega})+[1-\frac{a}{b}]\alpha}\right)^2}\right)\right\},\ \forall\omega\in[-\pi,\pi] \tag{4.121e}$$

$$V^{\star}_{i}(\mathrm{e}^{j\omega}) \triangleq \min\left\{1\ ,\ \frac{\frac{1}{2}(a/b)\alpha}{\sqrt{\lambda_i(\mathrm{e}^{j\omega})}\left(\sqrt{\lambda_i(\mathrm{e}^{j\omega})}+\sqrt{\lambda_i(\mathrm{e}^{j\omega})+[\frac{1}{a}-\frac{1}{b}]\frac{4}{\lambda}}\right)}\right\}.\ \forall\omega\in[-\pi,\pi] \tag{4.121f}$$

*Proof.* Let us define the stationary vector process

$$\mathbf{v}(k) \triangleq \boldsymbol{Q}(z)^H\mathbf{x}(k), \tag{4.122}$$

which has mutually independent elements, each of them having PSD $\lambda_i(\mathrm{e}^{j\omega})$, $i=1,\ldots,N$. Define also the vector processes

$$\mathbf{w}(k) \triangleq \boldsymbol{Q}(z)^H\mathbf{y}(k)$$
$$\mathbf{n}(k) \triangleq \mathbf{w}(k)-\mathbf{v}(k),$$

so that

$$\mathbf{y}(k)-\mathbf{x}(k) = \boldsymbol{Q}(z)\mathbf{n}(k). \tag{4.123}$$

Since $\boldsymbol{Q}(z)$ is unitary, we have that

$$D_{a,b}(\{\mathbf{x}(k)\},\{\mathbf{y}(k)\}) = D_{a,b}(\{\mathbf{v}(k)\},\{\mathbf{w}(k)\}),$$
$$\bar{I}(\{\mathbf{x}(k)\};\{\mathbf{y}(k)\}) = \bar{I}(\{\mathbf{v}(k)\};\{\mathbf{w}(k)\}),$$

and therefore the WCMSE-RDFs for $\{\mathbf{x}(k)\}$ and $\{\mathbf{v}(k)\}$ are equal.

On the other hand, since the processes in $\{\mathbf{v}(k)\}$ are mutually independent, it follows that, for every $\ell>1$, the matrix of eigenvectors of $\boldsymbol{K}_{\overline{\mathbf{v}}(\ell)}$ has block-diagonal structure. Hence, from Theorem 4.6, the vector $\overline{\mathbf{n}}(\ell)$ achieving $R_{a,b}(D)$ for $\overline{\mathbf{v}}(\ell)$ is Gaussian and is such that $\boldsymbol{K}_{\overline{\mathbf{n}}(\ell),\overline{\mathbf{v}}(\ell)}$ and $\boldsymbol{K}_{\overline{\mathbf{n}}(\ell)}$ are block-diagonal. This means that in order to achieve $R_{a,b}(D)$, $\mathrm{n}_i(k)$ must be independent of $\mathrm{n}_j(k)$ and of $\mathrm{v}_j(k)$, for every $k\in\mathbb{Z}$ and for every $j\neq i\in\mathbb{Z}$. Therefore, when $R_{a,b}(D)$ is achieved, the following holds

$$\bar{I}(\{\mathbf{v}(k)\};\{\mathbf{w}(k)\}) = \sum_{i=1}^{N}\bar{I}(\{\mathrm{v}_i(k)\};\{\mathrm{w}_i(k)\}), \tag{4.124}$$

see Lemma 4.4, and

$$D_{a,b}(\{\mathbf{v}(k)\},\{\mathbf{w}(k)\}) = \sum_{i=1}^{N}D_{a,b}(\{\mathrm{v}_i(k)\},\{\mathrm{w}_i(k)\}). \tag{4.125}$$

Thus, $R_{a,b}(D)$ can be regarded as a rate distortion function of a product source with a sum distortion metric. Since, in addition, the scalar processes $\{\mathrm{v}_i(k)\}$ are mutually independent, we can apply Theorem 4.5, which yields that

$$R_{a,b}(D) = \sum_{i=1}^{N} R_{a,b}^{(i)}(d_i) \tag{4.126a}$$

$$D = \sum_{i=1}^{N} d_i, \tag{4.126b}$$

where the distortions $\{d_i\}$ must satisfy

$$\frac{\partial}{\partial d_i} R_{a,b}^{(i)}(d_i) - s = 0, \tag{4.127}$$

for some slope $s < 0$, if $\frac{\partial}{\partial d} R_{a,b}^{(i)}(d_i^{max}) < s$, where $d_i^{max} \triangleq \min\{d : R_{a,b}^{(i)}(d) = 0\}$, or else

$$d_i = d_i^{max}.$$

From (4.83),

$$R_{a,b}^{(i)}(d_i) = \frac{1}{2\pi} \int\limits_{g^2 \geq \frac{a^2}{b^2}\alpha/4} \ln\left(g(\omega) + \sqrt{g(\omega)^2 + [1-\tfrac{a}{b}]\alpha}\right) - \frac{1}{2}\ln\alpha \; d\omega, \tag{4.128a}$$

where $g(\omega)^2 \triangleq \lambda_i(\mathrm{e}^{j\omega})$, $\forall \omega \in [-\pi, \pi]$, and $\alpha > 0$ is the unique scalar parameter satisfying

$$d_i = \frac{1}{2\pi} \int\limits_{g^2 \geq \frac{a^2}{b^2}\alpha/4} \frac{a\alpha g(\omega)/2}{g(\omega) + \sqrt{g(\omega)^2 + [1-\frac{a}{b}]\alpha}} d\omega + \frac{1}{2\pi} \int\limits_{g^2 < \frac{a^2}{b^2}\alpha/4} b g(\omega)^2(\mathrm{e}^{j\omega}) d\omega. \tag{4.128b}$$

Differentiating (4.128a) with respect to $\alpha$, we obtain:

$$\begin{aligned}
\frac{\partial}{\partial \alpha} R_{a,b}^{(i)} &= \frac{1}{2\pi} \int\limits_{g^2 \geq \frac{a^2}{b^2}\alpha/4} \frac{\frac{[1-\frac{a}{b}]}{2\sqrt{g^2+[1-\frac{a}{b}]\alpha}}}{g + \sqrt{g^2 + [1-\frac{a}{b}]\alpha}} - \frac{1}{2\alpha} d\omega \\
&= \frac{1}{2\pi} \int\limits_{g^2 \geq \frac{a^2}{b^2}\alpha/4} \frac{[1-\frac{a}{b}]\alpha - \left(g + \sqrt{g^2 + [1-\frac{a}{b}]\alpha}\right)\sqrt{g^2 + [1-\frac{a}{b}]\alpha}}{2\alpha\left(g + \sqrt{g^2 + [1-\frac{a}{b}]\alpha}\right)\sqrt{g^2 + [1-\frac{a}{b}]\alpha}} d\omega \\
&= \frac{1}{2\pi} \int\limits_{g^2 \geq \frac{a^2}{b^2}\alpha/4} \frac{-g\left(g + \sqrt{g^2 + [1-\frac{a}{b}]\alpha}\right)}{2\alpha\left(g + \sqrt{g^2 + [1-\frac{a}{b}]\alpha}\right)\sqrt{g^2 + [1-\frac{a}{b}]\alpha}} d\omega \\
&= \frac{1}{2\pi} \int\limits_{g^2 \geq \frac{a^2}{b^2}\alpha/4} \frac{-g}{2\alpha\sqrt{g^2 + [1-\frac{a}{b}]\alpha}} d\omega.
\end{aligned}$$

On the other hand, differentiation of (4.128b) with respect to $\alpha$ yields

$$\begin{aligned}
\frac{\partial}{\partial\alpha}d_i &= \frac{a/2}{2\pi}\int\limits_{g^2\geq\frac{a^2}{b^2}\alpha/4}\frac{g\left(g+\sqrt{g^2+[1-\frac{a}{b}]\alpha}\right)-\alpha g\frac{[1-\frac{a}{b}]}{2\sqrt{g^2+[1-\frac{a}{b}]\alpha}}}{(g+\sqrt{g^2+[1-\frac{a}{b}]\alpha})^2}d\omega \\
&= \frac{a/2}{2\pi}\int\limits_{g^2\geq\frac{a^2}{b^2}\alpha/4}\frac{2\left(g+\sqrt{g^2+[1-\frac{a}{b}]\alpha}\right)\sqrt{g^2+[1-\frac{a}{b}]\alpha}-[1-\frac{a}{b}]\alpha}{2(g+\sqrt{g^2+[1-\frac{a}{b}]\alpha})^2\sqrt{g^2+[1-\frac{a}{b}]\alpha}}g\,d\omega \\
&= \frac{a/2}{2\pi}\int\limits_{g^2\geq\frac{a^2}{b^2}\alpha/4}\frac{2g^2+2g\sqrt{g^2+[1-\frac{a}{b}]\alpha}+[1-\frac{a}{b}]\alpha}{2(g+\sqrt{g^2+[1-\frac{a}{b}]\alpha})^2\sqrt{g^2+[1-\frac{a}{b}]\alpha}}g\,d\omega \\
&= \frac{a/2}{2\pi}\int\limits_{g^2\geq\frac{a^2}{b^2}\alpha/4}\frac{(g+\sqrt{g^2+[1-\frac{a}{b}]\alpha})^2}{2(g+\sqrt{g^2+[1-\frac{a}{b}]\alpha})^2\sqrt{g^2+[1-\frac{a}{b}]\alpha}}g\,d\omega \\
&= \frac{a/2}{2\pi}\int\limits_{g^2\geq\frac{a^2}{b^2}\alpha/4}\frac{g}{2\sqrt{g^2+[1-\frac{a}{b}]\alpha}}\,d\omega.
\end{aligned}$$

Therefore,

$$\frac{\partial R^{(i)}_{a,b}(d_i)}{\partial d_i} = \frac{\partial R^{(i)}_{a,b}}{\partial\alpha}\Big/\frac{\partial d_i}{\partial\alpha} = -\frac{2}{a\alpha}. \tag{4.129}$$

This result, together with (4.123), (4.128) and (4.126), leads directly to (4.121),completing the proof. □

## 4.7 Image Processing Example

The purpose of this section is threefold. First, it provides an example illustrating the applicability of the results already presented in this chapter to the encoding of digital images. Secondly, it allows the reader to literally see the meaning of the WCMSE rate-distortion function in a more tangible manner. Finally, it supports the claim made at the end of Section 4.3.2, that in lossy image compression applications the perceptual weight of source-parallel distortion is greater than that of source-uncorrelated distortion. With this aim in mind, the linear distortion and additive source-independent noise required to realize $R_{a,b}(D)$ were applied to two black and white images. The results are shown in Fig. 4.5 and 4.6. More specifically, for a fixed rate of $0.05$ bits/sample, fixing $a = 1$, and for several values of $b$, each image was distorted according to the following steps:

1. Obtain the 2-dimensional DFT of the image.
2. Regard the DFT coefficients as samples of the power spectral density (PSD) of a w.s.s. vector process.

3. Plug the DFT coefficients into the "water filling" formula (4.121a) in place for the source PSD. Find the "water level" parameter $\alpha$ numerically.

4. Use $\alpha$ to find, for each frequency component, the gains that are required to realize $R_{a,b}(D)$, which are given by one minus the right hand side of (4.121f). These gain values correspond to samples of the signal transfer function required to realize $R_{a,b}(D)$.

5. Multiply each DFT coefficient of the source by its corresponding gain. Apply the inverse 2-D DFT to the result. This yields the linearly distorted image.

6. In order to generate the source-parallel noise similar to that obtained by subtractively dithered uniform quantizers, create an image with i.i.d. pixels, uniformly distributed, and obtain its 2-D DFT coefficients.

7. Multiply each of these coefficients by the the squared root of the right hand side of (4.121e), using the value of $\alpha$ found above, so as to obtain the noise PSD of a realization of $R_{a,b}(D)$. Apply the inverse 2-D DFT to the result. This yields approximately a realization of the source uncorrelated distortion in the pixel domain.

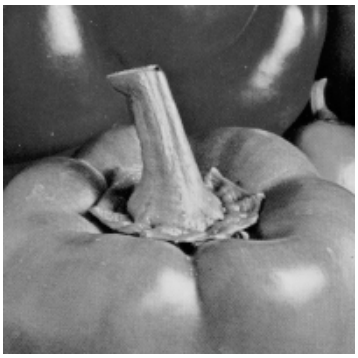
8. Add the source uncorrelated distortion to the linearly distorted image.

In each figure, the top left image corresponds to the original. We note again that all images were distorted using the same target bit-rate. As expected, images distorted using small values of $b$ suffer pronounced blurring and very mild additive noise. For the cases $a = b$, the result represents what is obtained when MSE is minimized under a constraint in the bit-rate. As the value of $b$ is increased, edges appear sharper and image contrast improves, at the expense of increased additive noise.

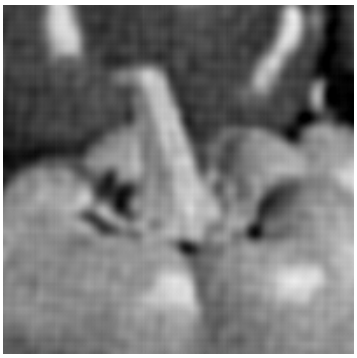
It is the opinion of the author and several other observers that in Figs. 4.5 and 4.6 the most acceptable images were distorted using a value of $b$ greater than $a$. This suggests that, for a fixed rate, the perceptual effect of linear distortion in near-optimal image compression is more objectionable than the effect of source uncorrelated noise.

## 4.8 Achievability

In this section, a proof of achievability for the WCMSE rate-distortion function is provided, for the case in which the source is a scalar Gaussian random process. The proof is based upon the use of optimal *entropy coded dithered lattice quantizers* (ECDLQ), which have been described in, e.g., [126]. We begin with a brief review of some of the main results related to these quantizers.



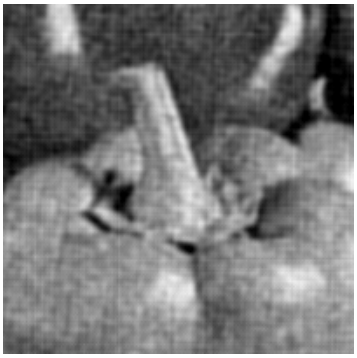
Original



$a = 1,\ b = 0.5$

$a = 1,\ b = 1$

$a = 1,\ b = 2$



$a = 1,\ b = 3$

$a = 1,\ b = 4$

$a = 1,\ b = 8$

$a = 1,\ b = 16$

$a = 1,\ b = 32$

**Figure 4.5:** Image linearly distorted plus filtered uniformly distributed and independent noise, in approximate accordance with $R_{a,b}(D)$ for several values of the parameter $b$ and for a fixed rate $R_{a,b} = 0.05$ bits/sample.


**Figure 4.6:** Image linearly distorted plus filtered uniformly distributed and independent noise, in approximate accordance with $R_{a,b}(D)$ for several values of the parameter $b$ and for a fixed rate $R_{a,b} = 0.05$ bits/sample.

### 4.8.1 Background on Dithered Lattice Quantization

A randomized lattice quantizer of dimension $N$ is an $N$-dimensional lattice quantizer $\mathcal{Q}_N$ with subtractive dither $\boldsymbol{\delta}$, followed by entropy encoding. The dither $\boldsymbol{\delta} \sim \mathcal{U}(\mathbb{V}_0)$ is uniformly distributed over a Voronoi cell $\mathbb{V}_0$ of the lattice quantizer. The idea of subtractive dither is to add the dither to the source prior to quantization, and then subtract the dither from the output of the quantizer to obtain the reconstruction vector. More precisely, the reconstructed output is given by

$$\mathbf{y} = \mathcal{Q}_N(\mathbf{x} + \boldsymbol{\delta}) - \boldsymbol{\delta}. \tag{4.130}$$

By doing this, the quantization error

$$\mathbf{e} = \mathbf{y} - \mathbf{x} = \mathcal{Q}_N(\mathbf{x} + \boldsymbol{\delta}) - \boldsymbol{\delta} - \mathbf{x}, \tag{4.131}$$

is distributed as $-\boldsymbol{\delta}$ and has covariance matrix

$$\boldsymbol{K}_{\mathbf{e}} = \epsilon \cdot \boldsymbol{I} \tag{4.132}$$

More importantly, $\mathbf{e}$ is independent of $\mathbf{x}$. Furthermore, it has been shown in [126] that the coding rate of the quantizer, i.e.

$$R_{\mathcal{Q}_N} \triangleq \tfrac{1}{N} H(\mathcal{Q}_N(\mathbf{x} + \boldsymbol{\delta}) | \boldsymbol{\delta}) \tag{4.133}$$

can be written as the mutual information between the source $\mathbf{x}$ and its reconstruction $\mathbf{y}$. More precisely,

$$R_{\mathcal{Q}_N} = \frac{1}{N} I(\mathbf{x}; \mathbf{y}) = \frac{1}{N} I(\mathbf{x}; \mathbf{x} + \mathbf{e}),$$

and the quadratic distortion per dimension is given by $\frac{1}{N}\mathrm{E}\left[\|\mathbf{y} - \mathbf{x}\|^2\right] = \frac{1}{N}\mathrm{E}\left[\|\mathbf{e}\|^2\right] = \frac{\epsilon}{N}$, see (4.132).

It has furthermore been shown that when $\boldsymbol{\delta}$ is white, there exists a sequence of lattice quantizers $\{\mathcal{Q}_N\}$ where the quantization error (and therefore also the dither) tends to be approximately Gaussian distributed (in the Kullback-Leibler divergence sense) for large $N$. Specifically, let $\mathbf{e}$ have PDF $f_{\mathbf{e}}(\cdot)$, and let $\mathbf{e}_G$ be Gaussian distributed with the same mean and covariance as $\mathbf{e}$. Then $\lim_{N\to\infty} \frac{1}{N} D(f_{\mathbf{e}} \| f_{\mathbf{e}_G}) = 0$ with a convergence rate of $\frac{\log(N)}{N}$, if the sequence $\{\mathcal{Q}_N\}$ is chosen appropriately [161].

In the next section we will be interested in the case where the dither is not necessarily white. By shaping the Voronoi cells of a lattice quantizer whose dither $\boldsymbol{\delta}$ is white, we also shape $\boldsymbol{\delta}$, obtaining a coloured dither $\boldsymbol{\delta}'$. This situation was considered in detail in [161] from where we obtain the following lemma (which was proven in [161] but not formally stated in the form of a lemma).

**Lemma 4.9.** *Let $\mathbf{e} \sim \mathcal{U}(\mathbb{V}_0)$ be white, i.e. $\mathbf{e}$ is uniformly distributed over the Voronoi cell $\mathbb{V}_0$ of the lattice quantizer $\mathcal{Q}_N$ and $\boldsymbol{K}_{\mathbf{e}} = \epsilon \boldsymbol{I}$. Furthermore, let $\mathbf{e}' \sim \mathcal{U}(\mathbb{V}_0')$, where $\mathbb{V}_0'$ denotes the shaped Voronoi cell*

*$\mathbb{V}_0' = \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{M}^{-1}\boldsymbol{x} \in \mathbb{V}_0\}$ and $\boldsymbol{M}$ is some invertible linear transformation. Denote the covariance of $\mathbf{e}'$ by $\boldsymbol{K}_{\mathbf{e}'} = \boldsymbol{M}\boldsymbol{M}^T\epsilon$. Similarly, let $\mathbf{e}_G \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}_{\mathbf{e}_G})$ having covariance matrix $\boldsymbol{K}_{\mathbf{e}_G} = \boldsymbol{K}_{\mathbf{e}}$ and let $\mathbf{e}_G' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}_{\mathbf{e}_G'})$ where $\boldsymbol{K}_{\mathbf{e}_G'} = \boldsymbol{K}_{\mathbf{e}'}$. Then there exists a sequence of shaped lattice quantizers such that*

$$\frac{1}{N} D(f_{\mathbf{e}'} \| f_{\mathbf{e}_G'}) = \mathcal{O}\left(\log(N)/N\right). \tag{4.134}$$

▲

*Proof.* The divergence is invariant to invertible transformations since $h(\mathbf{e}') = h(\mathbf{e}) + \log_2(|\boldsymbol{M}|)$. Thus, $D(f_{\mathbf{e}'} \| f_{\mathbf{e}_G'}) = D(f_{\boldsymbol{M}\mathbf{e}} \| f_{\boldsymbol{M}\mathbf{e}_G}) = D(f_{\mathbf{e}} \| f_{\mathbf{e}_G})$ for any $N$. □

### 4.8.2 Achievability of $R_{a,b}(D)$

In order to establish the achievability of $R_{a,b}(D)$, the following lemma is needed:

**Lemma 4.10.** Let $\mathbf{x}$, $\mathbf{x}'$, $\mathbf{z}$ and $\mathbf{z}'$ be mutually independent random vectors. Let $\mathbf{x}'$ and $\mathbf{z}'$ be arbitrarily distributed, and let $\mathbf{x}$ and $\mathbf{z}$ be Gaussian having the same mean and covariance as $\mathbf{x}'$ and $\mathbf{z}'$, respectively. Then

$$I(\mathbf{x}'; \mathbf{x}' + \mathbf{z}') \leq I(\mathbf{x}; \mathbf{x} + \mathbf{z}) + D(\mathbf{z}' \| \mathbf{z}). \tag{4.135}$$

▲

*Proof.*

$$\begin{aligned} I(\mathbf{x}'; \mathbf{x}' + \mathbf{z}') &= h(\mathbf{x}' + \mathbf{z}') - h(\mathbf{z}') \\ &\overset{(a)}{=} h(\mathbf{x} + \mathbf{z}) - h(\mathbf{z}) + D(\mathbf{z}' \| \mathbf{z}) - D(\mathbf{x}' + \mathbf{z}' \| \mathbf{x} + \mathbf{z}) \\ &\leq I(\mathbf{x}; \mathbf{x} + \mathbf{z}) + D(\mathbf{z}' \| \mathbf{z}), \end{aligned} \tag{4.136}$$

where $(a)$ stems from the well known result $D(\mathbf{x}' \| \mathbf{x}) = h(\mathbf{x}) - h(\mathbf{x}')$, see, e.g., [63, p. 254]. □

**Remark 4.5.** *For a uniform scalar subtractively dithered quantizer $\mathcal{Q}$, the net quantization noise $\mathrm{e}$ is uniformly distributed. In this case, if $\mathrm{e}_G \sim \mathcal{N}(0, \sigma_{\mathrm{e}}^2)$, then $D(\mathrm{e} \,\|\, \mathrm{e}_G) = 0.254$ bits. Hence, the scalar entropy of the quantized output conditioned on the dither exceeds the* scalar *mutual information obtained when $\mathcal{Q}$ is replaced by a channel with Gaussian AWN of variance $\sigma_{\mathrm{e}}^2$ by $0.254$ bits.* ▲

We can now prove the achievability of $R_{a,b}(D)$.

**Theorem 4.11.** *[Achievability of $R_{a,b}(D)$] Let $\mathbf{x}$ be an infinite length Gaussian random vector with zero mean. Define the sequence of vectors*

$$\mathbf{x}^{(N)} \triangleq \left[\mathrm{x}(-\tfrac{N-1}{2}), \mathrm{x}(-\tfrac{N-1}{2}+1), \ldots, \mathrm{x}(\tfrac{N-1}{2})\right]^T, \quad N \in \{2j+1\}_{j=0}^{\infty} \tag{4.137}$$

*Denote the covariance matrix of $\mathbf{x}^{(N)}$ by $\boldsymbol{K}_{\mathbf{x}^{(N)}}$. If*

$$\lim_{N\to\infty} \max\left\{\lambda_i(\boldsymbol{K}_{\mathbf{x}^{(N)}})\right\} < \infty, \tag{4.138}$$

*then $R_{a,b}(D)$ is achievable for all $D > 0$.* ▲

*Proof.* Let the eigenvalue decomposition of $\boldsymbol{K}_{\mathbf{x}^{(N)}}$ be

$$\boldsymbol{K}_{\mathbf{x}^{(N)}} = \boldsymbol{Q}_N \operatorname{diag}\left\{\sigma^2_{N,k}\right\} \boldsymbol{Q}_N^T, \quad k = 1, 2, \ldots, N. \tag{4.139}$$

From Theorem 4.6, the WCMSE-rate-distortion function for $\mathbf{x}^{(N)}$, say $R^{(N)}_{a,b}(D)$, is achieved when the reconstructed vector $\mathbf{y}^{(N)}$ has the form

$$\mathbf{y}^{(N)} = \mathbf{u}^{(N)} + (\boldsymbol{I} - \boldsymbol{V}_N)\mathbf{x}^{(N)}, \tag{4.140}$$

where $\mathbf{u}^{(N)} \in \mathbb{R}^N$ is a zero mean, Gaussian random vector with covariance matrix satisfying (4.62a), and where the linear distortion matrix $\boldsymbol{V}_N \in \mathbb{R}^{N\times N}$ satisfies (4.62c), assuming the source is $\mathbf{x}^{(N)}$. These covariance matrix of $\mathbf{u}^{(N)}$ and the matrix $\boldsymbol{V}_N$ have eigenvalue decomposition

$$\boldsymbol{K}_{\mathbf{u}^{(N)}} = \boldsymbol{Q}_N \operatorname{diag}\left\{\eta_k^{(N)}\right\} \boldsymbol{Q}_N^T \tag{4.141}$$

$$\boldsymbol{V}_N = \boldsymbol{Q}_N \operatorname{diag}\left\{p_k^{(N)}\right\} \boldsymbol{Q}_N^T, \tag{4.142}$$

where

$$\begin{aligned} &p_k^{(N)} \le 1, \quad \forall k \in \{1, \ldots, N\} \\ &p_k^{(N)} = 1 \iff \eta_k^N = 0 \iff \sigma^2_{N,k} \le 4(a/b)^2\alpha^{(N)}, \quad \forall k \in \{1, \ldots, N\}, \end{aligned} \tag{4.143}$$

and where $\alpha^{(N)} > 0$ is the scalar that satisfies (4.61b) when the eigenvalues of the source are $\sigma^2_{N,k}$. Notice that, if the latter are fixed, then $\alpha^{(N)}$ is a function only of $D$. To make this dependence explicit, we write

$$\alpha^{(N)} = \alpha^{(N)}(D) \tag{4.144}$$

The realization of $R_{a,b}(D)$ described by (4.140) can also be accomplished as follows: We first multiply the source $\mathbf{x}^{(N)}$ by $\boldsymbol{Q}_N^T$, obtaining the random vector

$$\bar{\mathbf{x}}^{(N)} \triangleq \boldsymbol{Q}_N^T \mathbf{x}^{(N)}, \tag{4.145}$$

which has covariance matrix $\boldsymbol{K}_{\bar{\mathbf{x}}^{(N)}} = \mathrm{diag}\left\{\sigma^2_{N,k}\right\}$. Then, the $k$-th element of $\bar{\mathbf{x}}^{(N)}$ is multiplied by a scalar gain $1 - p_k^{(N)}$, for each $k \in \{1, \ldots, N\}$, yielding the random vector

$$\tilde{\mathbf{x}}^{(N)} \triangleq \mathrm{diag}\left\{1 - p_k^{(N)}\right\} \bar{\mathbf{x}}^{(N)}. \tag{4.146}$$

A Gaussian noise vector $\mathbf{e}_G^{(N)}$, independent of $\mathbf{x}^{(N)}$ and having covariance matrix $\boldsymbol{K}_{\mathbf{e}_G^{(N)}} = \mathrm{diag}\{\eta_k^{(N)}\}$, is added to $\tilde{\mathbf{x}}^{(N)}$, which yields

$$\tilde{\mathbf{y}}^{(N)} = \mathbf{e}_G^{(N)} + \tilde{\mathbf{x}}^{(N)}. \tag{4.147}$$

Finally, $\tilde{\mathbf{y}}^{(N)}$ is multiplied by the unitary matrix $\boldsymbol{Q}_N$, yielding the reconstruction $\mathbf{y}^{(N)}$. More precisely,

$$\begin{aligned}
\boldsymbol{Q}_N \tilde{\mathbf{y}}^{(N)} &= \boldsymbol{Q}_N \left[\mathbf{e}_G^{(N)} + \tilde{\mathbf{x}}^{(N)}\right] = \boldsymbol{Q}_N \left[\mathbf{e}_G^{(N)} + \mathrm{diag}\left\{1 - p_k^{(N)}\right\} \bar{\mathbf{x}}^{(N)}\right] && (4.148)\\
&= \boldsymbol{Q}_N \left[\mathbf{e}_G^{(N)} + \mathrm{diag}\left\{1 - p_k^{(N)}\right\} \boldsymbol{Q}_N^T \mathbf{x}^{(N)}\right] && (4.149)\\
&= \mathbf{u}^{(N)} + (\boldsymbol{I} - \boldsymbol{V}_N)\mathbf{x}^{(N)} = \boldsymbol{y}^{(N)}. && (4.150)
\end{aligned}$$

We also have that

$$R_{a,b}^{(N)}(D) = \bar{I}(\mathbf{x}^{(N)}; \mathbf{y}^{(N)}) = \bar{I}(\mathbf{x}^{(N)}; \tilde{\mathbf{y}}^{(N)}) = \bar{I}(\bar{\mathbf{x}}^{(N)}; \tilde{\mathbf{y}}^{(N)}) = \bar{I}(\tilde{\mathbf{x}}^{(N)}; \tilde{\mathbf{x}}^{(N)} + \mathbf{e}_G^{(N)}) \tag{4.151}$$

Now, instead of adding the Gaussian noise $\mathbf{e}_G^{(N)}$ to $\tilde{\mathbf{x}},^{(N)}$, we can apply an ECDLQ to quantize the non-zero elements of $\tilde{\mathbf{x}}^{(N)}$. Denote the number of elements in $\tilde{\mathbf{x}}^{(N)}$ having non-zero variance by the function $L(N, D)$ (see (4.143) and (4.144)). Then, the ECDLQ would have dimension $L \leq N$. If the cell of this ECDLQ is shaped so that the source-independent error introduced by it, namely $\mathbf{e}^{(N)}$, has the same covariance matrix as the vector formed by the $L$ non-zero-variance elements of $\mathbf{e}_G^{(N)}$, then the end-to-end WCMSE would be the same as the one obtained with $\mathbf{e}_G^{(N)}$, i.e., it would be equal to $D$. Denote this quantizer by $\mathcal{Q}_L$. From (4.133), the entropy rate of $\mathcal{Q}_L$ would be

$$R_{\mathcal{Q}_L}(D) \triangleq \frac{1}{N} H(\mathcal{Q}_L(\tilde{\mathbf{x}}_L^{(N)} + \boldsymbol{\delta}^L)|\boldsymbol{\delta}^L) = \bar{I}(\tilde{\mathbf{x}}^N; \tilde{\mathbf{x}}^{(N)} + \mathbf{e}^{(N)}) \tag{4.152}$$

where $\tilde{\mathbf{x}}_L^{(N)}$ is the vector formed by removing the zero-variance elements of $\tilde{\mathbf{x}}^{(N)}$. Subtracting (4.151) from (4.152), we have that

$$R_{\mathcal{Q}_L}(D) - R_{a,b}^{(N)}(D) = \bar{I}(\tilde{\mathbf{x}}^N; \tilde{\mathbf{x}}^{(N)} + \mathbf{e}^{(N)}) - \bar{I}(\tilde{\mathbf{x}}^N; \tilde{\mathbf{x}}^{(N)} + \mathbf{e}_G^{(N)}). \tag{4.153}$$

Applying Lemma 4.10 to the latter,

$$\frac{1}{N} H(\mathcal{Q}_L) - R_{a,b}^{(N)}(D^{(N)}) \leq D(\mathbf{e}^{(N)} \| \mathbf{e}_G^{(N)}) \tag{4.154}$$

If, for a given $D > 0$,

$$\lim_{N\to\infty} L(N, D) = \infty \tag{4.155}$$

holds, then $L \to \infty$ as $N \to \infty$, and thus from Lemma 4.9 we have that $D(\mathbf{e}^{(N)} \| \mathbf{e}_G^{(N)}) \to 0$ as $N \to \infty$. Therefore, $\lim_{N\to\infty} R_{\mathcal{Q}_L}(D) = R_{a,b}(D)$. Else, if for a given $D$ we have that $\lim_{N\to\infty} L(N, D) < \infty$, and if (4.138) holds, then $\bar{I}(\tilde{\mathbf{x}}^N; \tilde{\mathbf{x}}^{(N)} + \mathbf{e}^{(N)})$, and thus $\bar{I}(\tilde{\mathbf{x}}^N; \tilde{\mathbf{x}}^{(N)} + \mathbf{e}_G^{(N)})$, tends to zero as $N \to \infty$. In view of (4.152) and (4.151), this implies that $0 = \lim_{N\to\infty} R_{\mathcal{Q}_L}(D) = \lim_{N\to\infty} R_{a,b}^{(N)}(D) = R_{a,b}(D)$. This completes the proof. □

**Remark 4.6** (Achievability for other Gaussian sources)**.**

1. *For zero mean stationary Gaussian random sources, $R_{a,b}(D)$ is achieved by taking $\mathbf{x}$ in Theorem 4.11 to be the complete input process.*

2. *For vector processes, the achievability of $R_{a,b}(D)$ follows by building $\mathbf{x}$ in Theorem 4.11 from the concatenation of infinitely many consecutive source vectors.*

## 4.9 $R^{\perp}(D)$ Within Feedback Loops

In this section we extend the definition and results regarding $R^{\perp}(D)$ to cases where a feedback path exists between the reconstruction and the source. We restrict the analysis to stationary processes only. The situation under study is depicted in Fig. 4.7. In this setting, $\{\mathrm{z}(k)\}$ and $\{\mathrm{r}(k)\}$ are two random processes external to the loop. The transfer functions $G_1(z)$, $G_2(z)$, $G_3(z)$ are linear, causal, and such that

$$\overline{G}(z) \triangleq G_1(z)G_2(z)G_3(z), \tag{4.156}$$

i.e., $\overline{G}(z)$ has a delay of at least one sample.

The transfer functions $G_1(z)$, $G_2(z)$, $G_3(z)$ are *not necessarily stable*. More specifically, we only require $\overline{G}(z)$ to be such that $1/(1 + \overline{G}(z))$ is stable, i.e., such that the closed loop system is stable. Our motivation to consider possibly unstable transfer functions stems from the fact that one of the main applications of feedback is precisely the stabilization of open-loop unstable systems [69].

In this setting, we regard the process $\{\mathrm{x}(k)\}$ as the source, and the process $\{\mathrm{y}(k)\}$ is the reconstruction. Before we extend the definition of $R^{\perp}(D)$ for this scheme, we will need to adapt the notions of rate and distortion to feedback scenarios. This is done next.

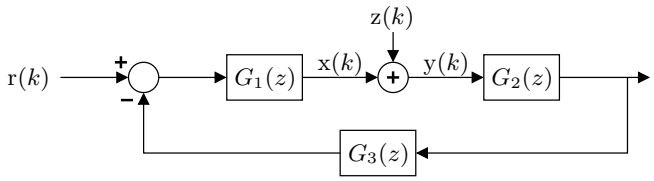



**Figure 4.7:** Source $\{\mathrm{x}(k)\}$ and reconstruction $\{\mathrm{y}(k)\}$ within the external feedback loop closed by the strictly causal transfer function $G_3(z)$.

### 4.9.1 The Directed Version of $R^{\perp}(D)$

The existence of a feedback path from $\{\mathrm{y}(k)\}$ to $\{\mathrm{x}(k)\}$ imposes the need to modify three fundamental aspects in the definition of $R^{\perp}(D)$:

1. the notion of mutual information to use;
2. the extent of the distortion un-correlation constraint; and
3. the signal whose variance is to be regarded as the distortion.

Each of these aspects is discussed below.

**Directed Mutual Information** When there is feedback from $\{\mathrm{y}(k)\}$ to $\{\mathrm{x}(k)\}$, the standard notion of mutual information needs to be replaced by that of *directed mutual information* [162]. In our case, this means that $R^{\perp}(D)$ needs to be redefined by using the *directed mutual information rate* $\bar{I}(\{\mathrm{x}(k)\} \to \{\mathrm{y}(k)\})$ instead of $\bar{I}(\{\mathrm{x}(k)\}\,;\{\mathrm{y}(k)\})$. For two random vectors $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^N$, the *directed mutual information* from $\mathbf{x}$ to $\mathbf{y}$ is defined as [162]:

$$I(\mathbf{x} \to \mathbf{y}) \triangleq \sum_{k=1}^{N} I(\mathrm{x}_1^k; \mathrm{y}(k) |\, \mathrm{y}^{k-1}) \tag{4.157}$$

For random processes, the above definition can be extended to the *directed mutual information rate*

$$\bar{I}(\{\mathrm{x}(k)\} \to \{\mathrm{y}(k)\}) \triangleq \lim_{\ell \to \infty} \frac{1}{\ell} \sum_{k=1}^{\ell} I(\mathrm{x}_1^k; \mathrm{y}(k) |\, \mathrm{y}_1^{k-1}). \tag{4.158}$$

If $\mathbf{y}$ depends causally on $\mathbf{x}$, and if there is no feedback from $\mathbf{y}$ to $\mathbf{x}$, then the following Markov chain holds:

$$\mathrm{x}_1^j \to \mathrm{x}_1^k \to \mathrm{y}_1^k, \quad 1 \le k \le j \le N. \tag{4.159}$$

When Markov chain (4.159) holds, we have that

$$I(\mathbf{x} \to \mathbf{y}) = I(\mathbf{x}; \mathbf{y}). \tag{4.160}$$

Consequently, mutual and directed mutual informations between $\mathbf{y}$ and $\mathbf{x}$ are equal if $\mathbf{y}$ depends causally on $\mathbf{x}$ and there exists no feedback from the former to the latter.

**Un-correlation Constraint.** The second fundamental aspect of $R^{\perp}(D)$ that needs to be modified whenever feedback is in place is the requirement of un-correlation between noise $\{\mathrm{z}(k)\}$ and source $\{\mathrm{x}(k)\}$. Assuming that $\{\mathrm{z}(k)\}$ is uncorrelated to $\{\mathrm{r}(k)\}$, and without feedback (i.e., if $G_3(z) \equiv 0$), one would have $\{\mathrm{x}(k)\}$ uncorrelated with $\{\mathrm{z}(k)\}$. On the contrary, when there is feedback, the processes $\{\mathrm{x}(k)\}$ and $\{\mathrm{z}(k)\}$ are not uncorrelated, since each element $\mathrm{x}(k)$ contains past samples of the process $\{\mathrm{y}(k)\} = \{\mathrm{x}(k)\} + \{\mathrm{z}(k)\}$. However, due to the linearity and (strict) causality of $\overline{G}(z)$, we have that if $\{\mathrm{r}(k)\}$ and $\{\mathrm{z}(k)\}$ are uncorrelated, then the $k$-th sample of the innovations process of $\{\mathrm{z}(k)\}$, namely $\mathrm{w}(k)$, *is indeed uncorrelated* to each of the samples $\{\mathrm{x}(i)\}_{i \leq k}$, i.e.,

$$\mathrm{E}\left[\mathrm{w}(k)\,\mathrm{x}(i)\right] = 0, \quad \forall i \in \mathbb{Z} : i \leq k, \forall k \in \mathbb{Z}. \tag{4.161}$$

We shall use this condition as the key constraint in our "feedback version" of $R^{\perp}(D)$, instead of requiring that $\{\mathrm{x}(k)\}$ and $\{\mathrm{z}(k)\}$ be uncorrelated.

**Weighted Quadratic Distortion.** In the general scheme depicted in Fig. 4.7, the forward channel $\{\mathrm{y}(k)\} = \{\mathrm{x}(k)\} + \{\mathrm{z}(k)\}$ forms part of a bigger system. For this reason, it makes sense to consider as a distortion metric the variance of, not only of $\{\mathrm{z}(k)\}$, but optionally, the variance of $\{\mathrm{z}(k)\}$ as it appears in other signals in the system, in the form of noise. This can be accomplished by considering the frequency weighting distortion metric $J(P(z), \{\mathrm{z}(k)\})$ defined in (4.109). In this case, the error weighting transfer function $P(z)$ represents the transfer function from $\{\mathrm{z}(k)\}$ to some given node in the system.

Based upon the above observations, we extend Definition 4.4 for the case of stationary random processes with feedback, as follows:

**Definition 4.6.** *In relation to the channel with feedback shown in Fig. 4.7, and for a given transfer function $P(z)$, we define the* ***quadratic rate-distortion function with source-uncorrelated distortion innovations*** *as*

$$\overrightarrow{R^{\perp}}(D) \triangleq \min_{\substack{\{\mathrm{z}(k)\}: J(P(z),\{\mathrm{z}(k)\}) \leq D, \\ E[\mathrm{x}(k)\,\mathrm{w}(j)]=0, \forall j \geq k \in \mathbb{Z}}} \bar{I}(\{\mathrm{x}(k)\} \to \{\mathrm{x}(k) + \mathrm{z}(k)\}), \quad D > 0, \tag{4.162}$$

*where $\{\mathrm{w}(k)\}$ is the innovations process underlying $\{\mathrm{z}(k)\}$, and where the frequency-weighted distortion metric $J(P(z), \{\mathrm{z}(k)\})$ is as in* (4.109). ▲

### 4.9.2 The Gaussian Case

In this section we characterize $R^{\overrightarrow{\perp}}(D)$ for the cases where $\{\mathrm{r}(k)\}$ in Fig. 4.7 is a Gaussian stationary process. The requirement of finite-delay imposed by feedback precludes the possibility of quantizing infinitely long sequences of samples of $\mathrm{x}$. Hence, the achievability for this minimum information rate seems impossible (unless infinitely many feedback loops operate in parallel, which would allow one to use infinite-dimensional vector quantization without introducing delay). Nevertheless, it is possible to get at least as close as $0.254$ bits/sample of this minimum rate by using subtractively dithered scalar quantization, see Remark 4.5.

The analysis will be carried out on the system depicted in Fig. 4.8. This system is equivalent to the one shown in Fig. 4.7 in terms of the signal transfer functions between $\{\mathrm{r}(k)\}$, $\{\mathrm{x}(k)\}$ and $\{\mathrm{y}(k)\}$. In particular, it can be easily verified that

$$\mathrm{x}(k) = \frac{G_1}{1+\overline{G}(z)}\,\mathrm{r}(k) - \frac{\overline{G}(z)}{1+\overline{G}(z)}\,\mathrm{z}(k), \tag{4.163}$$

as before, where $\overline{G}(z)$ is as defined in (4.156). Notice also that between $\{\tilde{\mathrm{r}}(k)\}$ and $\{\mathrm{y}(k)\}$ extends a perfect-reconstruction noise-shaping system, which can be seen as a special case of the source coders studied in Chapter 3.

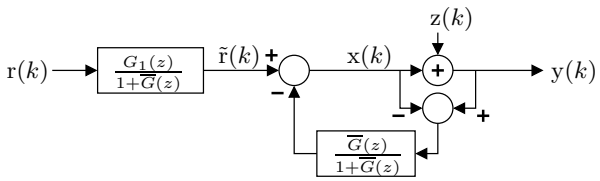



**Figure 4.8:** System equivalent, from $\{\mathrm{r}(k)\}$ to $\{\mathrm{x}(k)\}$ and $\{\mathrm{y}(k)\}$, to the one shown in Fig. 4.7. The transfer function $\overline{G}(z)$ corresponds to $G_1(z)G_2(z)G_3(z)$.

The following is a key theorem for subsequent results, and the first main result of this section. It states the relationship between the directed mutual information from $\{\mathrm{x}(k)\}$ to $\{\mathrm{y}(k)\}$ and the mutual information between $\{\tilde{\mathrm{r}}(k)\}$ and $\{\mathrm{y}(k)\}$. The main difficulty in the associated problem arises from the possible instability of $\overline{G}(z)$.

**Theorem 4.12** (Inner and Outer Information Rates Differ by the Entropy Gain of $(1+\overline{G}(z))^{-1}$)**.** *Consider the system depicted in Fig. 4.8, where $\{\tilde{\mathrm{r}}(k)\}$ and $\{\mathrm{z}(k)\}$ are random processes. Assume that the initial state of $\overline{G}(z)$ is a random vector having finite covariance matrix. If $\bar{I}(\{\tilde{\mathrm{r}}(k)\};\{\mathrm{y}(k)\})$ is*

*bounded, then*

$$\bar{I}(\{\mathrm{x}(k)\} \to \{\mathrm{y}(k)\}) \geq \bar{I}(\{\tilde{\mathrm{r}}(k)\}\,;\{\mathrm{y}(k)\}) + \sum_{p_i^{\overline{G}} \in \mathcal{P}} \log \left| p_i^{\overline{G}} \right|, \tag{4.164}$$

*where $\mathcal{P}$ is the set of unstable poles of $\overline{G}(z)$. Equality is achieved if $\{\tilde{\mathrm{r}}(k)\}$ and $\{\mathrm{z}(k)\}$ are independent.*

▲

*Proof.* For any integers $m, \ell \geq 1$, we have that

$$\begin{aligned}
&I(\mathrm{x}_1^{\ell} \to \mathrm{y}_1^{\ell}) - I(\tilde{\mathrm{r}}_1^{m}; \mathrm{y}_1^{m}) \\
&\qquad \stackrel{(a)}{=} I(\mathrm{x}_1^{\ell} \to \mathrm{y}_1^{\ell}) - I(\tilde{\mathrm{r}}_1^{m} \to \mathrm{y}_1^{m}) \\
&\qquad = \sum_{k=1}^{\ell} \left[ h(\mathrm{y}(k) | \, \mathrm{y}_1^{k-1}) - h(\mathrm{y}(k) | \, \mathrm{y}_1^{k-1}, \mathrm{x}_1^{k}) \right] - \sum_{j=1}^{m} \left[ h(\mathrm{y}(j) | \, \mathrm{y}_1^{j-1}) - h(\mathrm{y}(j) | \, \mathrm{y}_1^{j-1}, \tilde{\mathrm{r}}_1^{j}) \right] \\
&\qquad = \sum_{j=1}^{m} h(\mathrm{y}(j) | \, \mathrm{y}_1^{j-1}, \tilde{\mathrm{r}}_1^{j}) - \sum_{k=1}^{\ell} h(\mathrm{y}(k) | \, \mathrm{y}_1^{k-1}, \mathrm{x}_1^{k}) - h(\mathrm{y}_1^{m}) + h(\mathrm{y}_1^{\ell}),
\end{aligned} \tag{4.165}$$

where $(a)$ follows from the fact that there exists no feedback from $\{\mathrm{y}(k)\}$ to $\{\tilde{\mathrm{r}}(k)\}$. Define

$$\mathrm{n}(k) \triangleq \mathrm{y}(k) - \tilde{\mathrm{r}}(k), \quad \forall k \in \mathbb{Z}. \tag{4.166}$$

Notice that $\{\mathrm{n}(k)\}$ is completely (and causally) determined by $\{\mathrm{z}(k)\}$, since

$$\mathrm{n}(k) = \frac{1}{1 + \overline{G}(z)} \, \mathrm{z}(k) \tag{4.167}$$

and since $\frac{1}{1+\overline{G}(z)}$ is biproper. In addition, if $\{\mathrm{z}(k)\}$ depends on $\{\tilde{\mathrm{r}}(k)\}$, it does it causally, and without feedback from $\{\mathrm{z}(k)\}$ to $\{\tilde{\mathrm{r}}(k)\}$. Therefore, the following Markov chain holds:

$$\tilde{\mathrm{r}}_1^{\infty} \to \tilde{\mathrm{r}}_1^{k} \to \mathrm{z}_1^{k} \to \mathrm{n}_1^{k} \, . \tag{4.168}$$

We then have that, for every $m, \ell \geq 1$,

$$\begin{aligned}
&\sum_{j=1}^{m} h(\mathrm{y}(j)|\,\mathrm{y}_1^{j-1}, \tilde{\mathrm{r}}_1^{j}) - \sum_{k=1}^{\ell} h(\mathrm{y}(k)|\,\mathrm{y}_1^{k-1}, \mathrm{x}_1^{k}) \\
&\qquad \overset{(a)}{=} \sum_{j=1}^{m} h(\mathrm{n}(j)|\,\mathrm{y}_1^{j-1}, \tilde{\mathrm{r}}_1^{j}) - \sum_{k=1}^{\ell} h(\mathrm{z}(k)|\,\mathrm{y}_1^{k-1}, \mathrm{x}_1^{k}) \\
&\qquad \overset{(b)}{=} \sum_{j=1}^{m} h(\mathrm{n}(j)|\,\mathrm{n}_1^{j-1}, \tilde{\mathrm{r}}_1^{j}) - \sum_{k=1}^{\ell} h(\mathrm{z}(k)|\,\mathrm{z}_1^{k-1}, \mathrm{x}_1^{k}) \\
&\qquad \overset{(c)}{=} \sum_{j=1}^{m} h(\mathrm{n}(j)|\,\mathrm{n}_1^{j-1}, \tilde{\mathrm{r}}_1^{j}) - \sum_{k=1}^{\ell} h(\mathrm{z}(k)|\,\mathrm{z}_1^{k-1}, \tilde{\mathrm{r}}_1^{k}) \\
&\qquad \overset{(d)}{=} \sum_{j=1}^{m} h(\mathrm{n}(j)|\,\mathrm{n}_1^{j-1}, \tilde{\mathrm{r}}_1^{m}) - \sum_{k=1}^{\ell} h(\mathrm{z}(k)|\,\mathrm{z}_1^{k-1}, \tilde{\mathrm{r}}_1^{\ell}) \\
&\qquad \overset{(e)}{=} h(\mathrm{n}_1^{m}\,|\tilde{\mathrm{r}}_1^{m}) - h(\mathrm{z}_1^{\ell}\,|\tilde{\mathrm{r}}_1^{\ell}) \\
&\qquad \overset{(f)}{=} h(\mathrm{n}_1^{m}, \tilde{\mathrm{r}}_1^{m}) - h(\tilde{\mathrm{r}}_1^{m}) - h(\mathrm{z}_1^{\ell}, \tilde{\mathrm{r}}_1^{\ell}) + h(\tilde{\mathrm{r}}_1^{\ell}) \\
&\qquad \overset{(g)}{=} h(\tilde{\mathrm{r}}_1^{m}|\,\mathrm{n}_1^{m}) + h(\mathrm{n}_1^{m}) - h(\tilde{\mathrm{r}}_1^{m}) - h(\tilde{\mathrm{r}}_1^{\ell}|\,\mathrm{z}_1^{\ell}) - h(\mathrm{z}_1^{\ell}) + h(\tilde{\mathrm{r}}_1^{\ell}) \\
&\qquad \overset{(h)}{=} h(\mathrm{n}_1^{m}) - h(\mathrm{z}_1^{\ell}) + I(\tilde{\mathrm{r}}_1^{\ell}; \mathrm{z}_1^{\ell}) - I(\tilde{\mathrm{r}}_1^{m}; \mathrm{n}_1^{m})
\end{aligned} \tag{4.169}$$

In the above, $(a)$ and $(b)$ hold since $\mathrm{y}(k) = \mathrm{n}(k) + \tilde{\mathrm{r}}(k)$ and $\mathrm{y}(k) = \mathrm{x}(k) + \mathrm{z}(k)$. $(c)$ follows from the fact that, if $\mathrm{z}_1^{k-1}$ is known, then knowledge of $\mathrm{x}_1^{k}$ is equivalent to knowledge of $\tilde{\mathrm{r}}_1^{k}$ (see Fig. 4.8 and recall that $\frac{\overline{G}(z)}{1+\overline{G}(z)}$ is stable). $(d)$ stems from the Markov chain (4.168). $(e)$ stems from the chain rule of differential entropy (see Property 2.8 in Section 2.3). $(f)$, $(g)$ and $(h)$ follow from the property $h(a, b) = h(a|b) + h(b)$ and from the definition of mutual information, see (2.33). Substituting (4.169)

into (4.165), and using (4.158),

$$\begin{aligned}
&\bar{I}(\{\mathrm{x}(k)\} \to \{\mathrm{y}(k)\}) - \bar{I}(\{\tilde{\mathrm{r}}(k)\}\,;\,\{\mathrm{y}(k)\}) \\
&\qquad = \lim_{\ell\to\infty}\frac{1}{\ell} I(\mathrm{x}_1^{\ell} \to \mathrm{y}_1^{\ell}) - \lim_{m\to\infty}\frac{1}{m} I(\tilde{\mathrm{r}}_1^{m};\mathrm{y}_1^{m}) \\
&\qquad \overset{(a)}{=} \lim_{\ell\to\infty}\frac{1}{\ell} I(\mathrm{x}_1^{\ell} \to \mathrm{y}_1^{\ell}) - \lim_{m\to\infty}\frac{1}{m} I(\tilde{\mathrm{r}}_1^{m} \to \mathrm{y}_1^{m}) \qquad (4.170)\\
&\qquad \overset{(b)}{=} \lim_{m\to\infty}\frac{1}{m} h(\mathrm{n}_1^{m}) - \lim_{\ell\to\infty}\frac{1}{\ell} h(\mathrm{z}_1^{\ell}) + \lim_{\ell\to\infty}\frac{1}{\ell} I(\tilde{\mathrm{r}}_1^{\ell};\mathrm{z}_1^{\ell}) - \lim_{m\to\infty}\frac{1}{m} I(\tilde{\mathrm{r}}_1^{m};\mathrm{n}_1^{m}) \\
&\qquad\qquad + \lim_{m\to\infty}\frac{1}{m} h(\mathrm{y}_1^{m}) - \lim_{\ell\to\infty}\frac{1}{\ell} h(\mathrm{y}_1^{\ell}) \\
&\qquad \overset{(c)}{=} \bar{h}(\{\mathrm{n}(k)\}) - \bar{h}(\{\mathrm{z}(k)\}) + \lim_{\ell\to\infty}\frac{1}{\ell}\left[I(\tilde{\mathrm{r}}_1^{\ell};\mathrm{z}_1^{\ell}) - I(\tilde{\mathrm{r}}_1^{\ell};\mathrm{n}_1^{\ell})\right] \\
&\qquad \overset{(d)}{\geq} \bar{h}(\{\mathrm{n}(k)\}) - \bar{h}(\{\mathrm{z}(k)\}) \\
&\qquad \overset{(e)}{=} \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|\frac{1}{1+\overline{G}(\mathrm{e}^{j\omega})}\right|^2 d\omega = \sum_{p_i^{\overline{G}}\in\mathcal{P}} \log\left|p_i^{\overline{G}}\right|
\end{aligned}$$

where $(a)$ follows from the fact that there exists no feedback from $\{\mathrm{y}(k)\}$ to $\{\tilde{\mathrm{r}}(k)\}$, and thus mutual information and directed mutual information are equal. $(b)$ follows from substituting (4.169) into (4.165) and then into (4.170) $(c)$ follows from the definition of differential entropy rate, see (2.28). $(d)$ follows from the Markov chain (4.168) and from the Data-Processing Inequality, see Fact 2.5 on page 40. Notice that equality holds in $(d)$ if $\{\tilde{\mathrm{r}}(k)\}$ and $\{\mathrm{z}(k)\}$ are independent. $(e)$ follows from applying (2.27) to (4.167), and from the fact that $\frac{1}{1+\overline{G}(z)}$ is stable with a finite variance initial-state. The last equality follows from Jensen's formula [144] (see also the Bode Integral Theorem in, e.g., [145]). This completes the proof. □

The above theorem shows that the "internal" directed mutual information rate $\bar{I}(\{\mathrm{x}(k)\} \to \{\mathrm{y}(k)\})$ exceeds the "external" mutual information rate $\bar{I}(\{\tilde{\mathrm{r}}(k)\}\,;\{\mathrm{y}(k)\})$ by at least $\sum \log\left|p_i^{\overline{G}}\right|$. The latter is a non-negative quantity corresponding to the entropy gain of the transfer function $1/(1+\overline{G}(z))$. Notice also that Theorem 4.12 does not require $\{\tilde{\mathrm{r}}(k)\}$ or $\{\mathrm{z}(k)\}$ to be stationary, and that **it holds even if** $\{\tilde{\mathrm{r}}(k)\}$ **and** $\{\mathrm{z}(k)\}$ **are correlated**.

The next lemma will be useful to prove the second main result of this section:

**Lemma 4.13** (Mean Power Gain of Stable Filters for Non-Stationary Processes)**.** *Assume that*

$$\hat{\lambda}_{\mathrm{z}} \triangleq \max_{\ell}\max_{i}\left|\lambda_i\left(\boldsymbol{K}_{\mathrm{z}}^{(\ell)}\right)\right| < \infty.$$

*If two stable filters $P(z)$ and $\overline{P}(z)$ satisfy*

$$\left|P(\mathrm{e}^{j\omega})\right|^2 = \left|\overline{P}(\mathrm{e}^{j\omega})\right|^2 \leq M, \quad \textit{a.e. on } [-\pi,\pi], \qquad (4.171)$$

*for some bounded constant $M$, then*

$$J(P(z),\{\mathrm{z}(k)\}) = J(\overline{P}(z),\{\mathrm{z}(k)\}). \tag{4.172}$$

▲

*Proof.* The cost measure $J(P,\{\mathrm{z}(k)\})$ can be written in matrix form as follows:

$$J(P(z),\{\mathrm{z}(k)\}) \triangleq \lim_{\ell\to\infty}\frac{1}{\ell}\,\mathrm{tr}\left\{K_{\mathrm{z}}^{(\ell)}\boldsymbol{P}_\ell^H\boldsymbol{P}_\ell\right\} \tag{4.173}$$

where $\boldsymbol{P}_\ell$ is the $\ell\times\ell$ lower triangular Toeplitz matrix having the impulse response of $P(z)$ (truncated to the first $\ell$ samples) along its first column. Using this expression, the absolute value of the difference between the left and right hand sides of (4.172) can be bounded as

$$\begin{aligned}
&\left|J(P(z),\{\mathrm{z}(k)\}) - J(\overline{P}(z),\{\mathrm{z}(k)\})\right| = \left|\lim_{\ell\to\infty}\frac{1}{\ell}\,\mathrm{tr}\left\{K_{\mathrm{z}}^{(\ell)}\boldsymbol{P}_\ell^H\boldsymbol{P}_\ell\right\} - \lim_{\ell\to\infty}\frac{1}{\ell}\,\mathrm{tr}\left\{K_{\mathrm{z}}^{(\ell)}\overline{\boldsymbol{P}}_\ell^H\overline{\boldsymbol{P}}_\ell\right\}\right| \\
&= \lim_{\ell\to\infty}\frac{1}{\ell}\left|\mathrm{tr}\left\{K_{\mathrm{z}}^{(\ell)}\left(\boldsymbol{P}_\ell^H\boldsymbol{P}_\ell - \overline{\boldsymbol{P}}_\ell^H\overline{\boldsymbol{P}}_\ell\right)\right\}\right| \\
&\overset{(a)}{\leq} \hat{\lambda}_{\mathrm{z}}\lim_{\ell\to\infty}\frac{1}{\ell}\left|\mathrm{tr}\left\{\boldsymbol{P}_\ell^H\boldsymbol{P}_\ell - \overline{\boldsymbol{P}}_\ell^H\overline{\boldsymbol{P}}_\ell\right\}\right| \\
&\leq \hat{\lambda}_{\mathrm{z}}\lim_{\ell\to\infty}\left|\boldsymbol{P}_\ell^H\boldsymbol{P}_\ell - \overline{\boldsymbol{P}}_\ell^H\overline{\boldsymbol{P}}_\ell\right|_{HS},
\end{aligned} \tag{4.174}$$

where $|\cdot|_{HS}$ denotes the weak matrix norm, see Definition 2.5 on page 34. Inequality $(a)$ follows from Corollary 4.17. The last inequality in (4.174) follows from applying Jensen's inequality to (2.5).

In order to show that the last limit in (4.174) is zero, we will demonstrate that $\boldsymbol{P}_\ell^H\boldsymbol{P}_\ell$ and $\overline{\boldsymbol{P}}_\ell^H\overline{\boldsymbol{P}}_\ell$ are asymptotically equivalent (see Definition 2.7).

For this purpose, define

$$f_1(\omega) \triangleq P(\mathrm{e}^{j\omega}), \quad \forall\omega : \left|P(\mathrm{e}^{j\omega})\right| \leq M, \tag{4.175a}$$

$$f_2(\omega) \triangleq \overline{P}(\mathrm{e}^{j\omega}), \quad \forall\omega : \left|\overline{P}(\mathrm{e}^{j\omega})\right| \leq M, \tag{4.175b}$$

with $f_1(\omega_0) = M$, for all $\omega_0\in[-\pi,\pi]$ such that $\left|P(\mathrm{e}^{j\omega_0})\right| > M$, and $f_2(\omega_0) = M$, for all $\omega_0\in[-\pi,\pi]$ such that $\left|\overline{P}(\mathrm{e}^{j\omega_0})\right| > M$. Notice from (4.171) that

$$\left|f_1(\omega)\right|^2 = \left|f_2(\omega)\right|^2, \quad \text{a.e. on } [-\pi,\pi]. \tag{4.176}$$

The matrices $\boldsymbol{P}_\ell$ and $\overline{\boldsymbol{P}}_\ell$ can be written as

$$\boldsymbol{P}_\ell = \boldsymbol{T}_\ell(f_1), \tag{4.177}$$

$$\overline{\boldsymbol{P}}_\ell = \boldsymbol{T}_\ell(f_2), \tag{4.178}$$

where $\boldsymbol{T}_\ell(f_1)$ and $\boldsymbol{T}_\ell(f_2)$ are Wiener class Toeplitz matrices specified by $f_1$ and $f_2$, according to Definition 2.6. From the functions $f_1$ and $f_2$, also define the sequences of circulant matrices $\{\boldsymbol{C}_\ell(f_1)\}_{\ell=1}^{\infty}$ and $\{\boldsymbol{C}_\ell(f_2)\}_{\ell=1}^{\infty}$, according to Definition 2.6 (see Section 2.2). Since $P(z)$ and $\overline{P}(z)$ are stable, their associated impulse responses are absolutely summable. This, together with the fact that $f_1$ and $f_2$ are bounded by $M$, implies that the sequences of matrices $\{\boldsymbol{T}_\ell(f_1)\}_{\ell=1}^{\infty}$, $\{\boldsymbol{T}_\ell(f_2)\}_{\ell=1}^{\infty}$, $\{\boldsymbol{C}_\ell(f_1)\}_{\ell=1}^{\infty}$ and $\{\boldsymbol{C}_\ell(f_2)\}_{\ell=1}^{\infty}$ are uniformly bounded in the strong norm (see Definition 2.4). Hence, from [129, Lemma 11] the following asymptotic equivalences hold:

$$\boldsymbol{T}_\ell(f_1) \sim \boldsymbol{C}_\ell(f_1), \tag{4.179}$$

$$\boldsymbol{T}_\ell(f_2) \sim \boldsymbol{C}_\ell(f_2). \tag{4.180}$$

Direct application of [129, Theorem 1] yields

$$\boldsymbol{T}_\ell(f_1)^H \boldsymbol{T}_\ell(f_1) \sim \boldsymbol{C}_\ell(f_1)^H \boldsymbol{C}_\ell(f_1) \tag{4.181}$$

$$\boldsymbol{T}_\ell(f_2)^H \boldsymbol{T}_\ell(f_2) \sim \boldsymbol{C}_\ell(f_2)^H \boldsymbol{C}_\ell(f_2). \tag{4.182}$$

On the other hand, from (4.175), (4.171), and using [129, Lemma 10], we obtain

$$\boldsymbol{C}_\ell(f_1)^H \boldsymbol{C}_\ell(f_1) = \boldsymbol{C}_\ell(f_1^* f_1) = \boldsymbol{C}_\ell(f_2^* f_2) = \boldsymbol{C}_\ell(f_2)^H \boldsymbol{C}_\ell(f_2), \tag{4.183}$$

which gives

$$\boldsymbol{T}_\ell(f_1)\boldsymbol{T}_\ell(f_1)^H \sim \boldsymbol{C}_\ell(f_1 f_1^H) = \boldsymbol{C}_\ell(f_2 f_2^H) \sim \boldsymbol{T}_\ell(f_2)\boldsymbol{T}_\ell(f_2)^H. \tag{4.184}$$

By virtue of [129, Theorem 1], (4.184) leads directly to

$$\boldsymbol{T}_\ell(f_1)\boldsymbol{T}_\ell(f_1)^H \sim \boldsymbol{T}_\ell(f_2)\boldsymbol{T}_\ell(f_2)^H \tag{4.185}$$

which, by definition, implies that the limit on the right hand of (4.174) is zero. This completes the proof. □

Based upon Theorem 4.12 and Lemma 4.13, we can now state the second main result of this section.

**Theorem 4.14** ($\overrightarrow{R^{\perp}}(D)$ for Gaussian Stationary Loop-External Signals)**.** *Consider the closed-loop system shown in Fig. 4.7. If* $\{\mathrm{r}(k)\}$ *is a Gaussian stationary process, then*[6]

$$\overrightarrow{R^{\perp}}(D) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\frac{\sqrt{|G_1 P|^2 S_\mathrm{r} + \alpha} + |G_1 P|\sqrt{S_\mathrm{r}}}{\sqrt{\alpha}}\right) d\omega + \sum\nolimits_i \log\left|p_i^G\right|, \tag{4.186}$$

[6] In these expressions the argument $\mathrm{e}^{j\omega}$ of the functions in the integrands has been omitted for clarity.

*where $P(z)$ is a frequency weighting transfer function and where $\alpha > 0$ is the unique scalar satisfying*

$$D = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left( \sqrt{|G_1 P|^2 S_{\mathrm{r}} + \alpha} - |G_1 P| \sqrt{S_{\mathrm{r}}} \right) |G_1 P| \sqrt{S_{\mathrm{r}}} \, d\omega. \tag{4.187}$$

*Moreover, $\bar{I}(\{\mathrm{x}(k)\}; \{\mathrm{r}(k) + \mathrm{z}(k)\})$ equals $R^{\overrightarrow{\perp}}(D)$ if and only if $\{\mathrm{z}(k)\}$ is Gaussian, stationary, independent of $\{\mathrm{r}(k)\}$, and has PSD*

$$S_{\mathrm{z}}^{\star}(\mathrm{e}^{j\omega}) = \frac{1}{2} \frac{\left( \sqrt{|G_1 P|^2 S_{\mathrm{r}} + \alpha} - |G_1 P| \sqrt{S_{\mathrm{r}}} \right) |G_1| \sqrt{S_{\mathrm{r}}}}{|P|}, \quad \textit{a.e. on } [-\pi, \pi]. \tag{4.188}$$

▲

*Proof.* Define the error process

$$\mathrm{n}(k) \triangleq \mathrm{y}(k) - \tilde{\mathrm{r}}(k) = \frac{1}{1 + \overline{G}(z)} \, \mathrm{z}(k). \tag{4.189}$$

Let $H(z)$ be a bi-proper, stable, transfer function, having a stable inverse, and such that

$$\left| H(\mathrm{e}^{j\omega}) \right| = \left| 1 + \overline{G}(\mathrm{e}^{j\omega}) \right| \left| P(\mathrm{e}^{j\omega}) \right|, \quad \forall \omega \in [-\pi, \pi], \tag{4.190}$$

where $P(\mathrm{e}^{j\omega})$ is the error weighting frequency response of the frequency weighting filter $P(z)$ associated to $R^{\overrightarrow{\perp}}(D)$. From (4.112) and (4.113), if $\{\mathrm{n}(k)\}$ is uncorrelated with $\{\tilde{\mathrm{r}}(k)\}$, then the minimum of $\bar{I}(\{\tilde{\mathrm{r}}(k)\}; \{\tilde{\mathrm{r}}(k) + \mathrm{n}(k)\}$ subject to the constraint $J(H(z), \{\mathrm{n}(k)\}) \leq D$ is given by

$$\min_{\substack{\{\mathrm{n}(k)\}: \{\mathrm{n}(k)\} \perp \{\tilde{\mathrm{r}}(k)\} \\ J(H(z), \{\mathrm{n}(k)\}) \leq D}} \bar{I}(\{\tilde{\mathrm{r}}(k)\}; \{\tilde{\mathrm{r}}(k) + \mathrm{n}(k)\}) \tag{4.191}$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left( \frac{\sqrt{|H(\mathrm{e}^{j\omega})|^2 S_{\tilde{\mathrm{r}}}(\mathrm{e}^{j\omega})} + \left| H(\mathrm{e}^{j\omega}) \right| \sqrt{S_{\tilde{\mathrm{r}}}(\mathrm{e}^{j\omega}) + \alpha}}{\sqrt{\alpha}} \right) d\omega, \tag{4.192}$$

where $\alpha > 0$ is the unique scalar parameter satisfying

$$D = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\alpha \left| H(\mathrm{e}^{j\omega}) \right| \sqrt{S_{\tilde{\mathrm{r}}}(\mathrm{e}^{j\omega})}}{|H(\mathrm{e}^{j\omega})| \sqrt{S_{\tilde{\mathrm{r}}}(\mathrm{e}^{j\omega})} + \sqrt{|H(\mathrm{e}^{j\omega})|^2 S_{\tilde{\mathrm{r}}}(\mathrm{e}^{j\omega}) + \alpha}} d\omega. \tag{4.193}$$

From (4.114), the minimum in (4.192) is achieved if and only if the error $\{\mathrm{n}(k)\}$ is a Gaussian stationary process, independent of $\{\tilde{\mathrm{r}}(k)\}$, and having PSD

$$S_{\mathrm{n}}(\mathrm{e}^{j\omega}) = \frac{1}{2} \left( \sqrt{|H(\mathrm{e}^{j\omega})|^2 S_{\tilde{\mathrm{r}}}(\mathrm{e}^{j\omega}) + \alpha} - \left| H(\mathrm{e}^{j\omega}) \right| \sqrt{S_{\tilde{\mathrm{r}}}(\mathrm{e}^{j\omega})} \right) \frac{\sqrt{S_{\tilde{\mathrm{r}}}(\mathrm{e}^{j\omega})}}{|H(\mathrm{e}^{j\omega})|}, \quad \text{a.e. on } [-\pi, \pi]. \tag{4.194}$$

On the other hand, from (4.189),

$$H(z)\,\mathrm{n}(k) = \frac{H(z)}{1+\overline{G}(z)}\,\mathrm{z}(k) = \overline{P}(z)\,\mathrm{z}(k), \tag{4.195}$$

where

$$\overline{P}(z) \triangleq \frac{H(z)}{1+\overline{G}(z)}. \tag{4.196}$$

Notice from (4.190) and (4.196) that

$$\left|\overline{P}(\mathrm{e}^{j\omega})\right| = \left|P(\mathrm{e}^{j\omega})\right|, \quad \forall\omega\in[-\pi,\pi]. \tag{4.197}$$

Thus, for any process $\{\mathrm{z}(k)\}$,

$$J(H(z),\{\mathrm{n}(k)\}) = J(\overline{P}(z),\{\mathrm{z}(k)\}) = J(P(z),\{\mathrm{z}(k)\}), \tag{4.198}$$

where the last equality follows from Lemma 4.13. In view of Theorem 4.12, (4.198) implies that

$$\begin{aligned}\min_{\substack{\{\mathrm{n}(k)\}:\{\mathrm{n}(k)\}\perp\{\tilde{\mathrm{r}}(k)\}\\ J(H(z),\{\mathrm{n}(k)\})\le D}} &\bar{I}(\{\tilde{\mathrm{r}}(k)\};\{\tilde{\mathrm{r}}(k)+\mathrm{n}(k)\})\\ &= \min_{\substack{\{\mathrm{z}(k)\}:\{\mathrm{z}(k)\}\perp\{\tilde{\mathrm{r}}(k)\}\\ J(P(z),\{\mathrm{z}(k)\})\le D}} \bar{I}(\{\mathrm{x}(k)\};\{\mathrm{x}(k)+\mathrm{z}(k)\}) - \sum\nolimits_i \log\left|p_i^G\right|.\end{aligned} \tag{4.199}$$

In addition,

$$S_{\mathrm{n}}(\mathrm{e}^{j\omega}) = \left|\frac{1}{1+\overline{G}(\mathrm{e}^{j\omega})}\right|^2 S_{\mathrm{z}}(\mathrm{e}^{j\omega}) = \frac{\left|P(\mathrm{e}^{j\omega})\right|^2}{\left|H(\mathrm{e}^{j\omega})\right|^2} S_{\mathrm{z}}(\mathrm{e}^{j\omega}), \quad \forall\omega\in[-\pi,\pi] \tag{4.200a}$$

and

$$S_{\tilde{\mathrm{r}}}(\mathrm{e}^{j\omega}) = \left|\frac{G_1(\mathrm{e}^{j\omega})}{1+\overline{G}(\mathrm{e}^{j\omega})}\right|^2 S_{\mathrm{r}}(\mathrm{e}^{j\omega}) = \frac{\left|G_1(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2}{\left|H(\mathrm{e}^{j\omega})\right|^2} S_{\mathrm{r}}(\mathrm{e}^{j\omega}), \quad \forall\omega\in[-\pi,\pi] \tag{4.200b}$$

Substitution of (4.200) and (4.192) into (4.199) yields (4.186). Since (4.198) holds, (4.187) follows from substituting (4.200) into (4.193). Finally, substituting (4.200) into (4.194), it follows that a process $\{\mathrm{z}(k)\}$ minimizes $\bar{I}(\{\mathrm{x}(k)\}\to\{\mathrm{x}(k)+\mathrm{z}(k)\})$ subject to $J(\{\mathrm{z}(k)\})\le D$ if and only if it is Gaussian, stationary, independent of $\{\mathrm{r}(k)\}$ and has the PSD given in (4.188). This completes the proof. □

The following example illustrates the applicability of the results obtained in this section for networked control problems.

**Example:** Figure 4.9-(a) shows a model of the closed-loop control system introduced in Fig. 1.3 (see Section 1.1). The controller $C(z)$ and the plant $G(z)$ may be unstable, but the closed loop system is

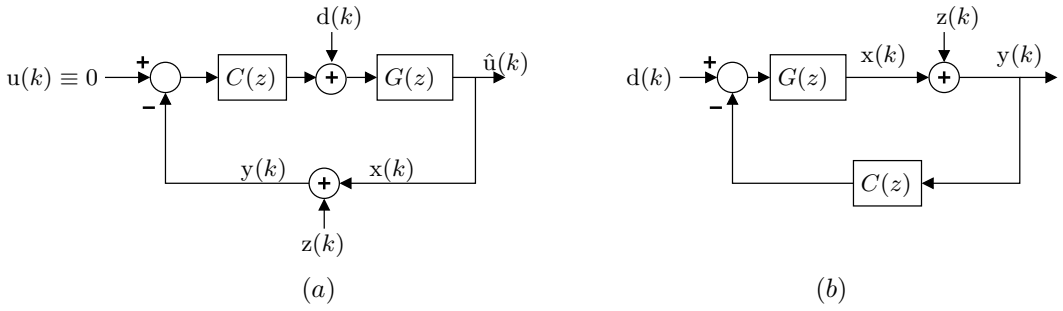



**Figure 4.9:** (a) Closed loop control system; (b) Equivalent scheme.

stable. The controller is bi-proper, and the plant has relative degree one or more. The reference signal $\{\mathrm{u}(k)\}$ is zero, and the disturbance $\{\mathrm{d}(k)\}$ is a Gaussian stationary process with PSD $S_d(\mathrm{e}^{j\omega})$. In Fig. 4.9-(a) the encoder-decoder pair has been replaced by a channel with additive noise $\{\mathrm{z}(k)\}$. This noise is the error introduced by the ED pair. An analytically equivalent system is shown in Fig. 4.9-(b). We are interested in finding the PSD of an error process $\{\mathrm{z}(k)\}$ uncorrelated to $\{\mathrm{d}(k)\}$, such that it minimizes the directed mutual information rate $\bar{I}(\{\mathrm{x}(k)\} \to \{\mathrm{y}(k)\})$ subject to the constraint that the variance of $\{\hat{\mathrm{u}}(k)\}$ is smaller than $D$. The variance of $\{\hat{\mathrm{u}}(k)\}$ is the tracking error variance. In order to apply Theorem 4.14, we need to determine the corresponding frequency weighing transfer function $P(z)$. Since we aim to minimize the variance of $\{\mathrm{x}(k)\}$, $P(z)$ needs to be the transfer function from $\{\mathrm{z}(k)\}$ to $\{\mathrm{x}(k)\}$. From Fig. 4.9, this transfer function is

$$P(z) = -\frac{C(z)G(z)}{1 + C(z)G(z)}. \tag{4.201}$$

The minimum of $\bar{I}(\{\mathrm{x}(k)\} \to \{\mathrm{y}(k)\})$, subject to having a tracking error variance smaller than $D$, for any $D > 0$, can be found directly by substituting the right-hand side of (4.201) for $P(z)$, $G(z)$ for $G_1(z)$, and $S_{\mathrm{d}}(\mathrm{e}^{j\omega})$ for $S_{\mathrm{r}}(\mathrm{e}^{j\omega})$, in Theorem 4.14.

## 4.10 Summary

In this chapter we have defined the rate-distortion function when the WCMSE is used as the distortion metric. We have characterized this RDF for Gaussian scalar sources, Gaussian vector sources, and Gaussian stationary process sources. The achievability of the WCMSE-RDF has been shown for Gaussian scalar and vector processes. We have seen that the WCMSE-RDF becomes Shannon's RDF for Gaussian sources when $a = b = 1$. It has also been verified that setting $a = 1$ and $b = \infty$ renders the WCMSE-RDF equivalent to the quadratic Gaussian RDF for source-uncorrelated distortions, denoted by $R^{\perp}(D)$,

recently introduced by the author in [127]. We have extended $R^{\perp}(D)$ to situations in which there exists linear, time-invariant feedback between reconstruction and source.

## 4.11 Appendix

**Lemma 4.15** (Theorem 4.5.2 in [6])**.** *Let $\boldsymbol{A}_{\infty}$ be an infinite Toeplitz matrix with entry $a_k \in \mathrm{r}$ on the $k$-th diagonal. Then the eigenvalues of $\boldsymbol{A}_{\infty}$ are contained in the interval $m \leq \lambda \leq M$, where $m$ and $M$ denote the essential infimum and supremum, respectively, of the function $f(\omega) \triangleq \sum_{k=-\infty}^{\infty} a_k \, \mathrm{e}^{-jk\omega}$. Moreover, if both $m$ and $M$ are finite and $G(\lambda)$ is any continuous function of $\lambda \in [m, M]$, then*

$$\lim_{N\to\infty} \frac{1}{N} \sum\nolimits_{k=1}^{N} G(\lambda_k^{(N)}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} G[f(\omega)] d\omega,$$

*where the $\lambda^{(N)}$ are the eigenvalues of the sub-matrix $\boldsymbol{A}^{(N)} \in \mathbb{R}^{N\times N}$ of $\boldsymbol{A}_{\infty}$ centred about the main diagonal of $\boldsymbol{A}_{\infty}$.* ▲

**Theorem 4.16** (From [163])**.** *If $\boldsymbol{A}$ and $\boldsymbol{B}$ are $N$-square normal matrices with eigenvalues $\lambda_i(\boldsymbol{A})$ and $\lambda_i(\boldsymbol{B})$, $i = 1, \ldots, N$, then*

$$\min \mathcal{R}\left\{\sum_{i=1}^{N} \lambda_i(\boldsymbol{A})\, \lambda_{p(i)}(\boldsymbol{B})\right\} \leq \mathcal{R}\left\{\sum_{i=1}^{N} \lambda_i(\boldsymbol{AB})\right\} \leq \max \mathcal{R}\left\{\sum_{i=1}^{N} \lambda_i(\boldsymbol{A})\, \lambda_{p(i)}(\boldsymbol{B})\right\}, \quad (4.202)$$

*where "$\max$" and "$\min$" are taken over all permutations $p$ of the eigenvalues of $\boldsymbol{B}$.* ▲

From this theorem, the next corollary follows immediately:

**Corollary 4.17.** *If $\boldsymbol{A}$ and $\boldsymbol{B}$ are $N$-square Hermitian matrices, with eigenvalues $\lambda_i(\boldsymbol{A})$ and $\lambda_i(\boldsymbol{B})$, $i = 1, \ldots, N$, where*

$$\lambda_i(\boldsymbol{A}) \geq \lambda_j(\boldsymbol{A}); \qquad \textit{and} \qquad \lambda_i(\boldsymbol{B}) \geq \lambda_j(\boldsymbol{B}), \qquad \forall i \geq j, \;\; i, j \in \{1, \ldots, N\}, \quad (4.203)$$

*then*

$$\min \sum_{i=1}^{N} \lambda_i(\boldsymbol{A})\, \lambda_{N+1-i}(\boldsymbol{B}) \leq \mathrm{tr}\,\{\boldsymbol{AB}\} \leq \max \sum_{i=1}^{N} \lambda_i(\boldsymbol{A})\, \lambda_i(\boldsymbol{B}), \quad (4.204)$$

*where "$\max$" and "$\min$" are taken over all permutations $p$ of the eigenvalues of $\boldsymbol{B}$.* ▲

# Chapter 5

# Using Realizations of the RDF to Design Optimal Source Coders

*The ten thousand questions are one question. If you cut through the one question, then the ten thousand questions disappear.*
*Zen proverb.*

## 5.1 Introduction

In this chapter we study how, and when, it is possible to utilize knowledge of a realization of the rate-distortion function, for a given source and distortion metric, to design optimal (or near-optimal) source coders. The source coders considered here comprise fullband and subband source coders based upon a quantizer and linear processing around it.

We will begin with an illustrative comparison. Consider the *feedback quantizer* (FQ) architecture shown in Fig. 3.1. We have seen that, given the Linear Model defined in Section 3.2.2, the *weighted correlation* MSE (WCMSE)-optimal filters in this scheme, under a constraint on the quantizer SNR, are characterized by (3.117). If $P(\mathrm{e}^{j\omega}) \equiv 1$, then the PSD of source-uncorrelated reconstruction errors obtained with these filters is

$$S_{\mathrm{u}}(\mathrm{e}^{j\omega}) \triangleq \sigma_{\mathrm{n}}^2 \left|B(\mathrm{e}^{j\omega})\right|^2 f(\mathrm{e}^{j\omega})^2 = \frac{\alpha}{4}\left(1 - \frac{\alpha}{\left[\sqrt{G(\mathrm{e}^{j\omega})^2 + [1-\frac{a}{b}]\alpha} + G(\mathrm{e}^{j\omega})\right]^2}\right), \quad \forall \omega \in [-\pi, \pi], \tag{5.1a}$$

where $\alpha > 0$ is the unique scalar satisfying (3.118). In addition, the frequency response of the signal

transfer function, given by (3.117b), is

$$A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega}) = 1 - \frac{(a/b)\alpha/2}{\left(\sqrt{G(\mathrm{e}^{j\omega})^2 + \left[1-\frac{a}{b}\right]\alpha} + G(\mathrm{e}^{j\omega})\right) G(\mathrm{e}^{j\omega})} \tag{5.1b}$$

By comparing (5.1) with (4.85), we see that

$$S_{\mathrm{u}}(\mathrm{e}^{j\omega}) = S_{\mathrm{u}}^{\star}(\mathrm{e}^{j\omega})$$

and

$$A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega}) = V^{\star}(\mathrm{e}^{j\omega}),$$

for all $\omega \in [-\pi, \pi]$. Therefore, (5.1a) and (5.1b) characterize a realization of $R_{a,b}(D)$ if source and quantization errors are replaced by Gaussian variables whilst keeping the same first and second moments. Equivalently, the SNR-optimal filters are also such that they minimize the mutual information between source and reconstruction, for a given value of the WCMSE.

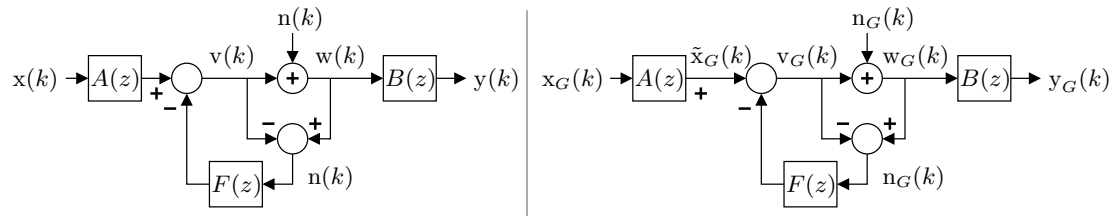
Clearly, if one knew beforehand that the optimal filters in an SNR-constrained optimization problem also realized $R_{a,b}(D)$, then it would have been possible to derive the optimal frequency responses obtained in Theorem 3.10 with ease. However, such correspondence does not always take place.

As an example, consider the case in which there is no feedback (i.e., when $F(z) \equiv 0$, see Fig. 3.1). The optimal frequency responses for $A(z)$ and $B(z)$ for this case (which are given by Theorem 3.5 taking $f(\mathrm{e}^{j\omega}) \equiv 1$), yield $S_{\mathrm{u}}(\mathrm{e}^{j\omega}) \neq S_{\mathrm{u}}^{\star}(\mathrm{e}^{j\omega})$ and $A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega}) \neq V^{\star}(\mathrm{e}^{j\omega})$, a.e. on $[-\pi, \pi]$, see (4.85a) and (4.85b). In other words, with the SNR-optimal filters characterized by Theorem 3.5, replacing source samples and quantization errors by Gaussian variables would not yield a realization of the WCMSE rate-distortion function (WCMSE-RDF).

The above comparisons raise the following questions:

- Why is SNR minimization at times, but not always, equivalent to end-to-end mutual information rate minimization?
- Why does the use of feedback in the first case examined above yield SNR minimizing filters that also realize the (WCMSE) rate-distortion function?
- Is it possible to know, a-priori, when such correspondence takes place in other schemes, such as, for example, subband coding architectures?

These are the main questions to be answered in this chapter. The answers will be given first for the case of scalar processes (which relates to FQ scheme discussed above), in Section 5.2, and then for the case of random vectors, in Section 5.3. In the latter case, we will see how to use knowledge of a




$\{\mathrm{x}(k)\}$ is w.s.s. with PSD $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$

$\{\mathrm{n}(k)\}$ is white and w.s.s.

$\mathrm{E}\,[\mathrm{x}(i)\,\mathrm{n}(k)] = 0,\ \forall i,k \in \mathbb{Z}$

$(a)$

$\{\mathrm{x}_G(k)\}$ is Gaussian stationary with PSD $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$

$\{\mathrm{n}_G(k)\}$ is white, Gaussian and stationary,
with variance $\sigma_{\mathrm{n}}^2$

$\mathrm{E}\,[\mathrm{x}_G(i)\,\mathrm{n}_G(k)] = 0,\ \forall i,k \in \mathbb{Z}$

$(b)$

**Figure 5.1:** a) Linear model of a scalar feedback quantizer. b) Forward test channel with filters.

realization of the WCMSE-RDF to design optimal transform coders. It will be shown that, under the Linear Model, the use of feedback is necessary for obtaining an optimal *causal* transform coder. In addition, we will show how to design optimal causal transform coders that are as rate-distortion efficient as the best non-causal transform coder, *at all rates*. Finally, we answer the above questions for the case of random vector processes in Section 5.4. We use the answers to characterize optimal *filter banks* (FBs), including the possible use of feedback. It is shown that, in general, and under the Linear Model, the use of all *three* degrees of freedom (pre-processing, post-processing and *feedback*) is necessary in order to obtain an optimal FB. By using the results derived in this section, it is possible to design FBs that attain an operational rate-distortion performance that exceeds the rate-distortion function by not more than $0.254$ bits/sample. Interestingly, under the Linear Model it turns out that, for optimality (which requires the use of feedback), the majorization property is not necessary. In particular, it is not necessary for optimality in perfect reconstruction filter banks. In addition, we showed that under the Linear Model, filter banks in which the subband signals (prior to quantization) are mutually uncorrelated are not optimal. These two observations stand in stark contrast with what is obtained for subband coders that do not make use feedback, see, e.g., [86, 113].

## 5.2 Conditions for Scalar Processes

It will be useful to formalize the questions stated at the end of the previous section by referring to the two systems depicted in Fig. 5.1, and their respective optimization problems, to be defined below. Notice that

the only difference between the schemes in Fig. 5.1-(a) and (b) is that, in the latter, the source $\{\mathrm{x}_G(k)\}$ and the noise $\{\mathrm{n}_G(k)\}$ are Gaussian. In particular, both systems share the same filters $A(z)$, $B(z)$ and $F(z)$, and equals SNRs:

$$\gamma = \frac{\sigma_{\mathrm{v}}^2}{\sigma_{\mathrm{n}}^2} = \frac{\sigma_{\mathrm{v}_G}^2}{\sigma_{\mathrm{n}_G}^2}, \tag{5.2}$$

since $S_{\mathrm{x}_G}(\mathrm{e}^{j\omega}) = S_{\mathrm{x}}(\mathrm{e}^{j\omega}),\ \forall\omega \in [-\pi, \pi]$, and $S_{\mathrm{n}_G}(\mathrm{e}^{j\omega}) = S_{\mathrm{n}}(\mathrm{e}^{j\omega}) = \sigma_{\mathrm{n}}^2,\ \forall\omega \in [-\pi, \pi]$.

Assuming that $\{\mathrm{n}(k)\}$ represents the sequence of quantization errors introduced by a scalar quantizer $\mathcal{Q}$, the scheme in Fig. 5.1-(a) can be regarded as an analysis model for feedback quantizers, as discussed in Chapter 3. More precisely, under the Linear Model, defined at the end of Section 3.2.2, quantization errors are white and uncorrelated with the source. In addition, if the output of $\mathcal{Q}$ is encoded in a memoryless fashion, then the associated operational bit-rate depends monotonically[1] on the SNR of $\mathcal{Q}$, i.e., on $\gamma$ (see Assumption 3.4 in Chapter 3). Therefore, under the Linear Model, the problem of minimizing the operational bit-rate under the constraint that the WCMSE does not exceed some value $D > 0$ can be stated as follows:

**Optimization Problem 5.1.** *In the scheme depicted in Fig. 5.1-(a),*

$$\textit{Minimize:}\ \ \gamma = \frac{\sigma_{\mathrm{v}}^2}{\sigma_{\mathrm{n}}^2} \tag{5.3}$$

$$\textit{Subject to:}\ \ D_{a,b}(\mathrm{x}, \mathrm{y}) \leq D \tag{5.4}$$

*over all filters $A(z)$, $B(z)$ and $F(z)$ such that the triplet $[A(z), B(z), F(z)] \in \mathbb{F}$, where $\mathbb{F}$ is a constraint set.* ▲

It must be noted that Optimization Problem 5.1 is the generalized converse of optimization problems 3.1–3.8 stated in Chapter 3. In each case, the architectural limitations that characterize each scenario are embodied in the constraint set $\mathbb{F}$.

The above optimization problem is the SNR minimization problem referred to in the questions at the end of Section 5.1. The end-to-end mutual information rate minimization problem in these questions can be formally defined with the help of the system depicted in Fig. 5.1-(b). This system can be utilized to obtain a forward test-channel realization of the WCMSE-RDF associated with the source $\{\mathrm{x}_G(k)\}$. In this configuration, the filters that yield such a realization must necessarily solve the following optimization problem (see Definition 4.2 in page 125):

[1] Strictly speaking, this statement is accurate only if one assumes that the changes in the PDF of v, stemming from varying $A(z)$, $B(z)$ and $F(z)$, have a negligible effect on the rate/SNR expressions (2.54) and (2.60). Nevertheless, the upper bound on the operational rate given in in (2.60) associated with subtractively dithered uniform scalar quantization is always valid.

**Optimization Problem 5.2.** *In the scheme depicted in Fig. 5.1-(b),*

$$\textit{Minimize:}\quad \bar{I}(\{\mathrm{x}_G(k)\}\,;\{\mathrm{y}_G(k)\}) \tag{5.5}$$

$$\textit{Subject to:}\quad D_{a,b}(\mathrm{x}_G,\mathrm{y}_G)\leq D \tag{5.6}$$

*over all filters $A(z)$, $B(z)$ and $F(z)$ such that $F(z)$ is strictly causal.* ▲

Notice that, in this optimization problem, there are no constraints on the filters $A(z)$, $B(z)$ and $F(z)$ (other than the strict causality of $F(z)$).

The following lemma states an important relationship between $\gamma$, $\bar{I}\left(\{\mathrm{x}_G(k)\}\,;\{\mathrm{y}_G(k)\}\right)$, and $R_{a,b}(D)$.

**Lemma 5.1.** *For the system in Fig. 5.1-(b), the following holds:*

$$\frac{\ln(\gamma+1)}{2}\overset{(a)}{=} I(\mathrm{v}_G(k);\mathrm{w}_G(k))\overset{(b)}{\geq}\bar{I}(\{\mathrm{v}_G(k)\}\to\{\mathrm{w}_G(k)\})\overset{(c)}{\geq}\bar{I}\left(\{\mathrm{x}_G(k)\}\,;\{\mathrm{y}_G(k)\}\right)\overset{(d)}{\geq}R_{a,b}(D). \tag{5.7}$$

*In addition,*

*i) Equality holds in $(b)$ if and only if $\{\mathrm{w}_G(k)\}$ is white, i.e., iff*

$$S_{\mathrm{w}_G}(\mathrm{e}^{j\omega})=\sigma^2_{\mathrm{w}_G}=\sigma^2_{\mathrm{w}},\qquad\forall\omega\in[-\pi,\pi]. \tag{5.8a}$$

*ii) Equality holds in $(c)$ if and only if*

$$\mathcal{N}_B\subseteq\mathcal{N}_A. \tag{5.8b}$$

*iii) Equality holds in $(d)$ if and only if*

$$\left|1-F(\mathrm{e}^{j\omega})\right|^2\left|B(\mathrm{e}^{j\omega})\right|\sigma^2_{\mathrm{n}_G}=S^\star_{\mathrm{u}}(\mathrm{e}^{j\omega}),\qquad\forall\omega\in[-\pi,\pi],\qquad\textit{and} \tag{5.8c}$$

$$A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega})=V^\star(\mathrm{e}^{j\omega}),\qquad\forall\omega\in[-\pi,\pi], \tag{5.8d}$$

*where $S^\star_{\mathrm{u}}(\mathrm{e}^{j\omega})$ and $V^\star(\mathrm{e}^{j\omega})$ are defined in* (4.85a) *and* (4.85b)*, respectively.*

▲

*Proof.* We proceed by parts.

- Equality $(a)$ follows from the fact that $\{\mathrm{v}_G(k)\}$ and $\{\mathrm{n}_G(k)\}$ are Gaussian and independent.

- Inequality $(b)$: We have that

$$\begin{aligned} I(\mathrm{v}_G(k);\mathrm{w}_G(k)) &= h(\mathrm{w}_G(k)) - h(\mathrm{w}_G(k)|\,\mathrm{v}_G(k)) = h(\mathrm{w}_G(k)) - h(\mathrm{v}_G(k)+\mathrm{n}_G\,|\,\mathrm{v}_G(k)) \\ &= h(\mathrm{w}_G(k)) - h(\mathrm{n}_G(k)|\,\mathrm{v}_G(k)) \\ &= h(\mathrm{w}_G(k)) - h(\mathrm{n}_G(k)) \qquad (5.9) \\ &= h(\mathrm{w}_G(k)) - h(\mathrm{n}_G(k)|\,\mathrm{n}_G^{k-1}) \qquad (5.10) \\ &\geq h(\mathrm{w}_G(k)|\,\mathrm{w}_G^{k-1}) - h(\mathrm{n}_G(k)|\,\mathrm{n}_G^{k-1}) \qquad (5.11) \\ &= I(\{\mathrm{v}_G(k)\} \to \{\mathrm{w}_G(k)\}) \end{aligned}$$

  In the above, (5.9) follows from the fact that $\{\mathrm{n}_G(k)\}$ and $\{\mathrm{x}_G(k)\}$ are independent and from the fact that $F(z)$ is strictly causal. As a consequence, $\mathrm{n}_G(k)$ is independent of $\mathrm{v}_G(k)$, for all $k \in \mathbb{Z}$. Similarly, (5.10) holds since the samples of $\{\mathrm{n}_G(k)\}$ are independent. Inequality (5.11) holds from the property $h(x|y) \leq h(x)$, with equality if and only if $x$ and $y$ are independent. This proves the first claim in Lemma 5.1.

- Inequality $(c)$: We have that

$$\begin{aligned} I(\{\mathrm{v}_G(k)\} \to \{\mathrm{w}_G(k)\}) &= \bar{h}(\{\mathrm{w}_G(k)\}) - h(\mathrm{w}_G(k)|\,\mathrm{w}_G^{k-1}, \mathrm{v}_G^k) \\ &= \bar{h}(\{\mathrm{w}_G(k)\}) - h(\mathrm{w}_G(k)|\,\mathrm{w}_G^{k-1}, \tilde{\mathrm{x}}_G^k) \qquad (5.12) \\ &= \bar{I}(\{\tilde{\mathrm{x}}_G(k)\} \to \{\mathrm{w}_G(k)\}) \\ &= \bar{I}(\{\tilde{\mathrm{x}}_G(k)\}\,;\{\mathrm{w}_G(k)\}) \qquad (5.13) \end{aligned}$$

  Equality in (5.12) holds from the fact that, if $\mathrm{w}_G^{k-1}$ is known, then $\tilde{\mathrm{x}}_G^k$ can be obtained deterministically from $\mathrm{v}_G^{k-1}$, and vice-versa. Equality (5.13) follows from the fact that there exists no feedback from $\{\mathrm{w}_G(k)\}$ to $\{\tilde{\mathrm{x}}_G(k)\}$. On the other hand, $\bar{I}(\{\tilde{\mathrm{x}}_G(k)\}\,;\{\mathrm{w}_G(k)\}) \geq \bar{I}(\{\mathrm{x}_G(k)\}\,;\{\mathrm{y}_G(k)\})$, with equality if and only if $B(z)$ is invertible for all frequencies $\omega$ for which $\left|A(\mathrm{e}^{j\omega})\right| > 0$.This proves the second claim in Lemma 5.1.

- Inequality $(d)$ follows from the definition of $R_{a,b}(D)$. The conditions for equality stated in point iii) in Lemma 5.1 follow directly from Theorem 4.7 on page 130.

This completes the proof. □

It is clear from (5.7) that, in the scheme depicted in Fig. 5.1-(b), the quantity $\frac{1}{2}\log(\gamma+1)$ is lower bounded by $R_{a,b}(D)$. Since there are no special constraints on the filters in Optimization Problem 5.2, it follows that condition iii) in Lemma 5.1 can always be met. Indeed, the combination of filters that solve Optimization Problem 5.2 (all of which yield $\bar{I}(\{\mathrm{x}_G(k)\}\,;\{\mathrm{y}_G(k)\}) = R_{a,b}(D)$) is not unique.

On the other hand, it is always possible to choose $A(z)$, $B(z)$ and $F(z)$ so that the *three* conditions in Lemma 5.1 are met. As a consequence, there exists, at least, one solution to Optimization Problem 5.2 for which $\frac{1}{2}\log(\gamma+1) = R_{a,b}(D)$.

These observations raise the following question: Is the WCMSE-RDF for $\{\mathrm{x}_G(k)\}$ also a lower bound for $\frac{1}{2}\log(\gamma+1)$ in the system shown in Fig. 5.1-(a)? The answer is yes, as shown by the following lemma:

**Lemma 5.2.** *In the system depicted in Fig. 5.1-(a), the following holds:*

$$\frac{1}{2}\log(\gamma+1) \geq R_{a,b}(D), \tag{5.14}$$

*where $R_{a,b}(D)$ is the WCMSE-RDF for the source $\{\mathrm{x}_G(k)\}$, which is Gaussian and has the same PSD as $\{\mathrm{x}(k)\}$. Equality is achieved if and only if conditions (i), (ii) and (iii) in Lemma 5.1 are met.* ▲

*Proof.* The validity of the result will be shown by using a contradiction argument. Thus, suppose that (5.14) does not hold. Then, there exist a triplet of filters in $\mathbb{F}$ such that $D_{a,b}(\mathrm{x},\mathrm{y}) \leq D$ and $\frac{1}{2}\log(\gamma+1) < R_{a,b}(D)$. If these filters are now used in the scheme of Fig. 5.1-(b), then the value of $\gamma$ would be the same. In addition, we have that $D_{a,b}(\mathrm{x}_G,\mathrm{y}_G) = D_{a,b}(\mathrm{x},\mathrm{y}) \leq D$ (since the WCMSE depends only on the second moments of the source and reconstruction). However, this contradicts (5.7), proving the validity of (5.14). □

Lemma 5.2 leads to the first main result of this section:

**Theorem 5.3.** *Suppose there exists a triplet of filters $[A(z), B(z), F(z)] \in \mathbb{F}$ that satisfies (5.8). Then, a triplet of filters $[A'(z), B'(z), F'(z)] \in \mathbb{F}$ is a solution to Optimization Problem 5.1 if and only if $[A'(z), B'(z), F'(z)]$ satisfies (5.8).* ▲

*Proof.* If $[A(z), B(z), F(z)] \in \mathbb{F}$ satisfies (5.8), then, from Lemma 5.1, it yields a value for $\gamma$ such that $\ln(\gamma+1)/2 = R_{a,b}(D)$. Thus, $[A(z), B(z), F(z)]$ yields the minimum $\gamma$ that can be achieved with any filters. Therefore, a triplet $[A'(z), B'(z), F'(z)] \in \mathbb{F}$ is a solution to Optimization Problem 5.1 only if it yields $\ln(\gamma+1)/2 = R_{a,b}(D)$. From Lemma 5.1, the latter holds if and only if $[A'(z), B'(z), F'(z)]$ satisfies (5.8). This completes the proof. □

Theorem 5.3 states an easy to verify condition under which filters that minimize the SNR $\gamma$, for a constraint $D_{a,b}(\mathrm{x},\mathrm{n}) < D$, would also realize $R_{a,b}(D)$ if the source and the noise were Gaussian. When these conditions are met, knowledge of the realization of the WCMSE-RDF can be used directly to determine the optimal filters in a scalar feedback quantizer under the Linear Model.

The next corollary follows immediately from Theorem 5.3:

**Corollary 5.4.** *Suppose there exists a triplet of filters* $[A(z), B(z), F(z)] \in \mathbb{F}$ *that satisfy* (5.8)*. Then, every solution to Optimization Problem 5.1 is also a solution to Optimization Problem 5.2.* ▲

Notice that if $A(z)$, $B(z)$ and $F(z)$ are unconstrained design choices, then (5.8) can be met for any $S_{\mathrm{u}}^{\star}$ and $V^{\star}(\mathrm{e}^{j\omega})$. In such cases, Corollary 5.4 implies that Optimization Problem 5.2 can be solved indirectly by solving Optimization Problem 5.1.

### 5.2.1 All Three Degrees of Freedom are Necessary

A key conclusion to be drawn from Lemma 5.2 is that, in the system of Fig. 5.1-(a), *at least three degrees of freedom are required in order to yield the ultimately achievable minimum for* $\gamma$. This stems from the fact that the frequency responses $A(\mathrm{e}^{j\omega})$, $B(\mathrm{e}^{j\omega})$ and $F(\mathrm{e}^{j\omega})$ need to satisfy the three equations (5.8a), (5.8c), and (5.8d), and from noting that

$$S_{\mathrm{w}}(\mathrm{e}^{j\omega}) = \left|A(\mathrm{e}^{j\omega})\right|^2 S_{\mathrm{x}}(\mathrm{e}^{j\omega}) + \sigma_{\mathrm{n}}^2 \left|1 - F(\mathrm{e}^{j\omega})\right|^2, \quad \forall \omega \in [-\pi, \pi]. \tag{5.15}$$

As we will see next, the use of entropy coding with memory allows one to obtain optimal performance with only two of the degrees of freedom embodied by $A(z)$, $B(z)$ and $F(z)$.

### 5.2.2 Entropy Coding with Memory is an Extra Degree of Freedom

Here we show that, when subtractively dithered uniform scalar quantization is employed in a feedback quantizer, then the use of entropy coding with infinite memory constitutes an additional degree of freedom, apart from the three provided by the filters around the quantizer. We restrict to the cases in which the source $\{\mathrm{x}(k)\}$ in Fig. 5.1-(a) is stationary.

In order to demonstrate the above claim, the following technical preliminary results are necessary.

#### Preliminary Results

The following result is the continuous analogue of that obtained by Kramer for discrete random variables [164, Property 3.6].

**Lemma 5.5.** *In Fig. 5.1-(a), let* $\{\mathrm{v}(k)\}$, $\{\mathrm{w}(k)\}$ *be jointly stationary random processes. If the differential entropy rates of* $\{\mathrm{w}(k)\}$ *and* $\{\mathrm{n}(k)\} \triangleq \{\mathrm{w}(k) - \mathrm{v}(k)\}$ *are bounded, then*

$$\bar{I}(\{\mathrm{v}(k)\} \to \{\mathrm{w}(k)\}) = \lim_{k \to \infty} I(\mathrm{v}_1^k; \mathrm{w}(k) | \, \mathrm{w}_1^{k-1}). \tag{5.16}$$

▲

*Proof.* We have that

$$\lim_{k\to\infty} I(\mathrm{v}_1^k; \mathrm{w}(k)|\, \mathrm{w}_1^{k-1}) \overset{(a)}{=} \lim_{i\to\infty} h(\mathrm{w}(k)|\, \mathrm{w}_1^{k-1}) - \lim_{k\to\infty} h(\mathrm{w}(k)|\, \mathrm{w}_1^{k-1}, \mathrm{v}_1^k) \tag{5.17a}$$

$$\overset{(b)}{=} \bar{h}(\mathrm{w}) - \lim_{k\to\infty} h(\mathrm{w}(k)|\, \mathrm{w}_1^{k-1}, \mathrm{v}_1^k). \tag{5.17b}$$

Equality $(a)$ holds if and only if the each of the limits on the right hand side of (5.17a) exist. The of these limits, $\lim_{i\to\infty} h(\mathrm{w}(0)|\, \mathrm{w}_{-i}^{-1})$, is by definition the entropy rate of $\{\mathrm{w}(k)\}$, which from the requirements of the lemma, exists. Thus, equality $(b)$ holds if and only if the second limit on the right hand side of (5.17b) exists. To show that such limit exists, we note that

$$h(\mathrm{w}(k)|\, \mathrm{w}_1^{k-1}, \mathrm{v}_1^k) = h(\mathrm{n}(k)|\, \mathrm{w}_1^{k-1}, \mathrm{v}_1^k) \geq h(\mathrm{n}(k)|\, \mathrm{n}_1^{k-1}), \quad \forall k \in \mathbb{Z}^+, \tag{5.18}$$

where the inequality follows from the Markov chain $(\mathrm{w}_1^{k-1}, \mathrm{v}_1^k) \leftrightarrow \mathrm{n}_1^{k-1} \leftrightarrow \mathrm{n}(k)$. The latter stems from the fact that the samples of $\{\mathrm{n}(k)\}$ are independent both mutually and with respect to $\{\mathrm{x}(k)\}$; and from fact that $\mathrm{w}_1^{k-1}$ and $\mathrm{v}_1^k$ are linear combinations of the samples of $\mathrm{n}_1^{k-1}$ and samples of $\{\mathrm{x}(k)\}$. Taking limits on both sides of (5.18) yields

$$\lim_{k\to\infty} h(\mathrm{w}(k)|\, \mathrm{w}_1^{k-1}, \mathrm{v}_1^k) \geq \bar{h}(\{\mathrm{n}(k)\}). \tag{5.19}$$

From the stationarity of $\{\mathrm{w}(k)\}$, it follows that $h(\mathrm{w}(k)|\, \mathrm{w}_1^{k-1}, \mathrm{v}_1^k)$ decreases monotonically with increasing $k$. This result, together with (5.19) and the fact that $\left|\bar{h}(\{\mathrm{n}(k)\})\right| < \infty$, implies that $\lim_{k\to\infty} h(\mathrm{w}(k)|\, \mathrm{w}_1^{k-1}, \mathrm{v}_1^k)$ exists. This proves that $(b)$ and $(c)$ in (5.17) hold, and that $\lim_{k\to\infty} I(\mathrm{v}_1^k; \mathrm{w}(k)|\, \mathrm{w}_1^{k-1})$ exists. The validity of (5.16) then follows directly by virtue of Cesáro mean theorem, see., e.g., [63, Thm. 4.2.3]. This completes the proof. □

**Lemma 5.6.** *In Fig. 5.1-(a), assume that $\{\mathrm{x}(k)\}$ is a stationary source, and that $\{\mathrm{n}(k)\}$ is i.i.d. noise introduced by a subtractively dithered uniform scalar quantizer (SDUSQ), $\mathcal{Q}$. Let the process $\{\mathrm{q}(k)\}$ denote the quantized output of $\mathcal{Q}$. Then*

$$\frac{1}{N} H(\mathrm{q}_{Nk-N+1}^{Nk}\,|\, \mathrm{q}_1^{Nk-N}, \delta_1^{Nk}) = \bar{I}(\{\mathrm{v}(k)\} \to \{\mathrm{w}(k)\}). \tag{5.20}$$

▲

*Proof.*

$$\begin{aligned}
\frac{1}{N}H(\mathrm{q}_{Nk-N+1}^{Nk}\,|\,\mathrm{q}_1^{Nk-N},\delta_1^{Nk}) &= \frac{1}{N}\sum_{i=N(k-1)+1}^{Nk} H(\mathrm{q}(i)|\,\mathrm{q}_1^{i-1},\delta_1^{Nk}) \\
&\overset{(a)}{=} \frac{1}{N}\sum_{i=N(k-1)+1}^{Nk} H(\mathrm{q}(i)|\,\mathrm{q}_1^{i-1},\delta_1^{i}) \\
&\overset{(b)}{=} \frac{1}{N}\sum_{i=N(k-1)+1}^{Nk} \left[H(\mathrm{q}(i)|\,\mathrm{q}_1^{i-1},\delta_1^{i}) - H(\mathrm{q}(i)|\,\mathrm{q}_1^{i-1},\delta_1^{i},\mathrm{v}_1^{i})\right] \\
&= \frac{1}{N}\sum_{i=N(k-1)+1}^{Nk} I\left(\mathrm{v}_1^{i};\mathrm{q}(i)|\,\mathrm{q}_1^{i-1},\delta_1^{k}\right) \\
&= \frac{1}{N}\sum_{i=N(k-1)+1}^{Nk} \left[h(\mathrm{v}_1^{i}\,|\,\mathrm{q}_1^{i-1},\delta_1^{i}) - h(\mathrm{v}_1^{i}\,|\,\mathrm{q}_1^{i},\delta_1^{i})\right] \\
&\overset{(c)}{=} \frac{1}{N}\sum_{i=N(k-1)+1}^{Nk} \left[h(\mathrm{v}_1^{i}\,|\,\mathrm{w}_1^{i-1},\delta(i)) - h(\mathrm{v}_1^{i}\,|\,\mathrm{w}_1^{i})\right] \\
&\overset{(d)}{=} \frac{1}{N}\sum_{i=N(k-1)+1}^{Nk} \left[h(\mathrm{v}_1^{i}\,|\,\mathrm{w}_1^{i-1}) - h(\mathrm{v}_1^{i}\,|\,\mathrm{w}_1^{i})\right] \\
&= \frac{1}{N}\sum_{i=N(k-1)+1}^{Nk} I(\mathrm{v}_1^{i};\mathrm{w}(i)|\,\mathrm{w}_1^{i-1}) \qquad (5.21)
\end{aligned}$$

In the above, $(a)$ follows from the fact that all dither samples in $\delta_{i+1}^{\infty}$ are independent of all samples $\mathrm{q}_1^i$, for all $i \in \mathbb{Z}^+$. Equality $(b)$ stems from the fact that $\mathrm{q}(i)$ is a deterministic function of $\mathrm{v}(i)$ and $\delta(i)$, which yields $H(\mathrm{q}(i)|\,\mathrm{q}_1^{i-1},\delta_1^i,\mathrm{v}_1^i) = 0$. Equality $(c)$ holds from the fact that knowledge of $\mathrm{w}(i)$ is equivalent to knowledge of $\{\mathrm{q}(i),\delta(i)\}$, $\forall i \in \mathbb{Z}^+$. The latter is a consequence of the fact that the reconstruction levels of $\mathcal{Q}$ are the midpoints of intervals of length $\Delta$, together with the fact that the dither in an SDUSQ satisfies $|\delta(i)| \in [-\frac{\Delta}{2},\frac{\Delta}{2})$, $\forall i$. Equality $(c)$ follows from the fact that, in and SDUSQ, the dither sample $\delta(i)$ is independent of $\mathrm{v}_1^i$ and independent of $\mathrm{w}_1^i$.

Taking the limit as $k \to \infty$, we obtain

$$\begin{aligned}
\lim_{k\to\infty}\frac{1}{N}H(\mathrm{q}_{Nk-N+1}^{Nk}\,|\,\mathrm{q}_1^{Nk-N},\delta_1^{Nk}) &= \lim_{k\to\infty}\frac{1}{N}\sum_{i=N(k-1)+1}^{Nk} I(\mathrm{v}_1^{i};\mathrm{w}(i)|\,\mathrm{w}_1^{i-1}) \\
&= \lim_{i\to\infty} I(\mathrm{v}_1^{i};\mathrm{w}(i)|\,\mathrm{w}_1^{i-1}) \qquad (5.22)
\end{aligned}$$

where (5.22) follows from the fact that $\{\mathrm{v}(k)\}$ and $\{\mathrm{w}(k)\}$ are jointly stationary. Finally, (5.20) follows directly from (5.22) upon applying Lemma 5.5. This completes the proof. □

**Theorem 5.7.** *In Fig. 5.1-(a), assume that $\{\mathrm{x}(k)\}$ is a stationary source, and that $\{\mathrm{n}(k)\}$ is i.i.d. noise introduced by a subtractively dithered uniform scalar quantizer (SDUSQ), $\mathcal{Q}$. Let $R_{op}^{N}$ be the minimum expected codeword length per symbol achievable when encoding $N$-length blocks $\mathrm{q}_{Nk-N+1}^{Nk}$ of the quantized output of $\mathcal{Q}$, when $k \to \infty$. Assume that both encoder and decoder have knowledge of dither samples $\delta_1^{Nk}$ and of past quantized outputs $\mathrm{q}_1^{Nk-N}$. Then*

$$\bar{I}(\{\mathrm{v}(k)\} \to \{\mathrm{w}(k)\}) \leq R_{op}^{N} \leq \bar{I}(\{\mathrm{v}(k)\} \to \{\mathrm{w}(k)\}) + \frac{1}{N}. \tag{5.23}$$

*In addition,*

$$R_{op}^{\infty} \triangleq \lim_{N\to\infty} R_{op}^{N} = \bar{I}(\{\mathrm{v}(k)\} \to \{\mathrm{w}(k)\}). \tag{5.24}$$

*Proof.* From [63, Theorem 5.4.2], it follows directly that

$$\lim_{k\to\infty} \frac{1}{N} H(\mathrm{q}_{Nk-N+1}^{Nk} \,|\, \mathrm{q}_1^{Nk-N}, \delta_1^{Nk}) \leq R_{op}^{N} \leq \lim_{k\to\infty} \frac{1}{N} H(\mathrm{q}_{Nk-N+1}^{Nk} \,|\, \mathrm{q}_1^{Nk-N}, \delta_1^{Nk}) + \frac{1}{N} \tag{5.25}$$

Substitution of (5.20) into the above yields (5.23), from which (5.24) follows immediately. This completes the proof. □

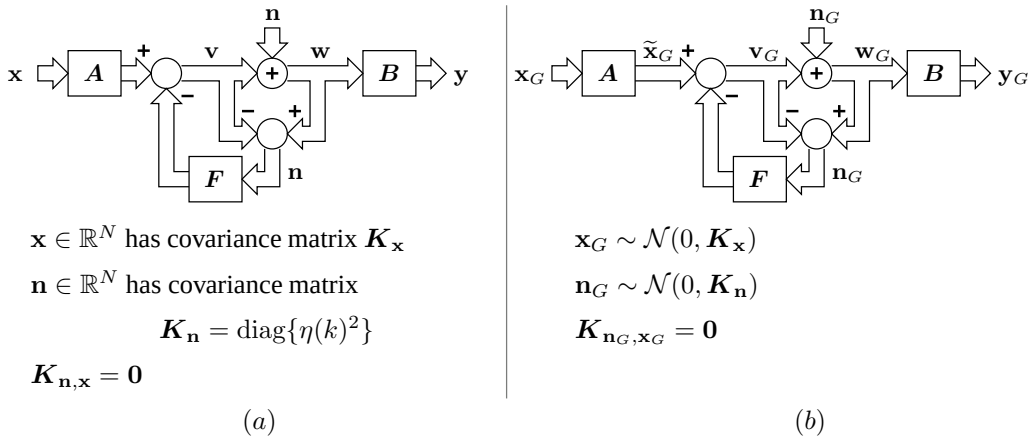
**Entropy Coding with Memory is an Extra Degree of Freedom**

From Theorem 5.7, and if $\{\mathrm{x}(k)\}$ is a Gaussian stationary source, it follows that the minimum achievable operational rate when using SDUSQ together with entropy coding with memory, $R_{op}^{\infty}$, in bits/sample, satisfies

$$\bar{I}(\{\mathrm{v}_G(k)\} \to \{\mathrm{w}_G(k)\}) \leq R_{op}^{\infty} \leq \bar{I}(\{\mathrm{v}_G(k)\} \to \{\mathrm{w}_G(k)\}) + 0.254. \tag{5.26}$$

This means that the minimal achievable operational bit-rate decouples from the SNR $\gamma$ and from the scalar mutual information $I(\mathrm{v}_G(k); \mathrm{w}_G(k))$. Only conditions (ii) and (iii) in Lemma 5.1 need to be met in order to minimize $\bar{I}(\{\mathrm{v}_G(k)\} \to \{\mathrm{w}_G(k)\})$. Therefore, if we associate the operational rate with the upper bound in (5.26), then *only the two equations* (5.8d) and (5.8c) need to be satisfied in order to minimize $R_{op}$. As a consequence, when SDUSQ is employed in a scalar feedback quantizer encoding a stationary source, the use of entropy coding with memory allows one to attain the minimal achievable operational bit-rate without any of the three degrees of freedom associated with the filters $A(z)$, $B(z)$ and $F(z)$ (see (5.8a), (5.8c) and (5.8d)).

Results similar to those obtained in this section,can be obtained for the cases where the source is a random vector, as discussed next.




$\mathbf{x} \in \mathbb{R}^N$ has covariance matrix $\boldsymbol{K}_{\mathbf{x}}$

$\mathbf{n} \in \mathbb{R}^N$ has covariance matrix

$$\boldsymbol{K}_{\mathbf{n}} = \mathrm{diag}\{\eta(k)^2\}$$

$\boldsymbol{K}_{\mathbf{n},\mathbf{x}} = \mathbf{0}$

$(a)$

$\mathbf{x}_G \sim \mathcal{N}(0, \boldsymbol{K}_{\mathbf{x}})$

$\mathbf{n}_G \sim \mathcal{N}(0, \boldsymbol{K}_{\mathbf{n}})$

$\boldsymbol{K}_{\mathbf{n}_G,\mathbf{x}_G} = \mathbf{0}$

$(b)$

**Figure 5.2:** a) Linear model of a transform coder. b) Forward vector test channel.

## 5.3 Conditions for Vector Sources

In this section we derive results analogue to Lemmas 5.1 and 5.2, Theorem 5.3, and Corollary 5.4, for the cases in which the source is an $N$-dimensional random vector. For this purpose, consider the system shown in Fig. 5.2-(a). In this figure, $\mathbf{x}$ and $\mathbf{n}$ are zero mean random vectors, and $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{F}$ are $N \times N$ matrices. The elements of the random vector $\mathbf{n}$ are mutually uncorrelated, and $\mathbf{n}$ is uncorrelated with $\mathbf{x}$. This system could be regarded as an analysis model for a transform coder with feedback [57, 120, 165]. Using the Linear Model, $\mathbf{n}$ would represent the error introduced by $N$ parallel scalar quantizers. In order to avoid algebraic loops, the matrix $\boldsymbol{F}$ needs to be lower triangular with zeros along its main diagonal (i.e., $\boldsymbol{F}$ needs to be *strictly causal*). This constraint is analogous to requiring the feedback filter $F(z)$ in the systems in Fig. 5.1 to be strictly causal.

Let us define the *vector of signal-to-noise ratios*

$$\boldsymbol{\gamma} \triangleq [\gamma(1), \gamma(2), \ldots, \gamma(N)]^T, \tag{5.27}$$

where

$$\gamma(k) \triangleq \frac{\sigma^2_{\mathrm{v}(k)}}{\eta(k)^2}, \quad k = 1, 2, \ldots, N \tag{5.28}$$

denotes the scalar SNR in the $k$-th channel, $\sigma^2_{\mathrm{v}(k)}$ is the variance of the $k$-th element of $\mathbf{v}$ and

$$\eta(k)^2 \triangleq \mathrm{E}\left[\mathrm{n}(k)^2\right], \quad k = 1, 2, \ldots, N, \tag{5.29}$$

is the variance of the $k$-th element in $\mathbf{n}$.

We define the following optimization problem:

**Optimization Problem 5.3.** *For the system depicted in Fig. 5.2-(a),*

$$\textit{Minimize: } \frac{1}{2N}\sum_{k=1}^{N}\log_2(\gamma(k)+1) \tag{5.30}$$

$$\textit{Subject to: } D_{a,b}(\mathbf{x},\mathbf{y}) \leq D, \tag{5.31}$$

*over all sets of non-negative noise variances* $\{\eta(k)^2\}_{k=1}^{N}$ *and over all matrices* $\boldsymbol{A}$, $\boldsymbol{B}$ *and* $\boldsymbol{F}$ *such that* $\boldsymbol{F}$ *is strictly causal and* $[\boldsymbol{A},\boldsymbol{B},\boldsymbol{F}] \in \mathbb{F}$*, where* $\mathbb{F}$ *is a constraint set of matrix triplets.* ▲

The second system, depicted in Fig. 5.2-(b), differs from that of Fig. 5.2-(a) only in that the source and the noise are the Gaussian random vectors $\mathbf{x}_G$ and $\mathbf{n}_G$, having the same covariance matrices as $\mathbf{x}$ and $\mathbf{n}$, respectively. Since both $\mathbf{x}_G$ and $\mathbf{n}_G$ are Gaussian, the un-correlation condition $\boldsymbol{K}_{\mathbf{n}_G,\mathbf{x}G} = \mathbf{0}$ implies that $\mathbf{x}_G$ and $\mathbf{n}_G$ are independent. Similarly the fact that $\boldsymbol{K}_{\mathbf{n}_G}$ is diagonal implies that $\mathbf{n}_G$ has mutually independent elements. This system can be seen as a forward vector channel realization of the WCMSE-RDF for vector Gaussian sources, as characterized in Section 4.4. From Definition 4.2, the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{F}$ that yield a realization of $R_{a,b}(D)$ for $\mathbf{x}_G$ must necessarily solve the following optimization problem:

**Optimization Problem 5.4.** *For the system depicted in Fig. 5.2-(b),*

$$\textit{Minimize: } \frac{1}{N}I(\mathbf{x}_G;\mathbf{y}_G) \tag{5.32}$$

$$\textit{Subject to: } D_{a,b}(\mathbf{x}_G,\mathbf{y}_G) \leq D, \tag{5.33}$$

*over all sets of non-negative noise variances* $\{\eta(k)^2\}_{k=1}^{N}$ *and over all square matrices* $\boldsymbol{A}$, $\boldsymbol{B}$ *and* $\boldsymbol{F}$ *such that* $\boldsymbol{F}$ *is strictly causal.* ▲

Clearly, the vectors of SNRs in both systems in Fig. 5.2 are the same, since, in both systems, corresponding signals have the same second moments. The following lemma, which is the vector version of Lemma 5.1, establishes a key relationship between $\boldsymbol{\gamma}$ and $R_{a,b}(D)$:

**Lemma 5.8.** *In the system depicted in Fig. 5.2-(b), the following holds:*

$$\frac{1}{2N}\sum_{k=1}^{N}\log(\gamma(k)+1) \overset{(a)}{=} \frac{1}{N}\sum_{k=1}^{N}I(\mathrm{v}_G(k);\mathrm{w}_G(k)) \overset{(b)}{\geq} \bar{I}(\mathbf{v}_G \to \mathbf{w}_G) \overset{(c)}{\geq} \bar{I}(\mathbf{x}_G;\mathbf{y}_G) \overset{(d)}{\geq} R_{a,b}(D) \tag{5.34}$$

*In addition,*

*i) Equality is achieved in* $(b)$ *if and only if* $\boldsymbol{K}_{\mathbf{w}_G}$ *is a diagonal matrix.*

*ii) Equality is achieved in* $(c)$ *if and only if* $\mathcal{N}_{\boldsymbol{B}} \subseteq \mathcal{R}_{\boldsymbol{A}}^{\perp}$.

*iii) Equality is achieved in* $(d)$ *iff*

$$\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{F})\boldsymbol{K}_{\mathbf{n}_G}(\boldsymbol{I}-\boldsymbol{F})^H\boldsymbol{B}^H = \boldsymbol{K}_{\mathbf{u}}^{\star} \tag{5.35a}$$

$$\boldsymbol{B}\boldsymbol{A} = \boldsymbol{I}-\boldsymbol{V}^{\star}, \tag{5.35b}$$

*where* $\boldsymbol{K}_{\mathbf{u}}^{\star}$ *and* $\boldsymbol{V}^{\star}$ *are as defined in* (4.62)*.*

▲

*Proof.* We proceed by parts.

- Equality $(a)$ follows from the fact that $\mathbf{x}_G$ and $\mathbf{n}_G$ are Gaussian and independent, together with the fact that $\boldsymbol{F}$ is strictly causal.

- Inequality $(b)$: We have that

$$\begin{aligned}
I(\mathrm{v}_G(k);\mathrm{w}_G(k)) &= h(\mathrm{w}_G(k)) - h(\mathrm{w}_G(k)|\,\mathrm{v}_G(k)) = h(\mathrm{w}_G(k)) - h(\mathrm{v}_G(k)+\mathrm{n}_G(k)|\,\mathrm{v}_G(k)) \\
&= h(\mathrm{w}_G(k)) - h(\mathrm{n}_G(k)|\,\mathrm{v}_G(k)) \\
&= h(\mathrm{w}_G(k)) - h(\mathrm{n}_G(k)) && (5.36)\\
&= h(\mathrm{w}_G(k)) - h(\mathrm{n}_G(k)|\mathrm{n}_G{}_1^{k-1}) && (5.37)\\
&\geq h(\mathrm{w}_G(k)|\mathrm{w}_G{}_1^{k-1}) - h(\mathrm{n}_G(k)|\mathrm{n}_G{}_1^{k-1}) && (5.38)\\
&\Rightarrow \frac{1}{N}\sum_{k=1}^{N} I(\mathrm{v}_G(k);\mathrm{w}_G(k)) \geq \bar{I}(\mathbf{v}_G \to \mathbf{w}_G)
\end{aligned}$$

  In the above, (5.36) follows from the fact that $\mathbf{n}_G$ and $\mathbf{x}_G$ are independent and that $F(z)$ is strictly causal. As a consequence, $\mathrm{n}_G(k)$ is independent of $\mathrm{v}_G(k)$, for all $k \in \{1,\ldots,N\}$. Similarly, (5.37) holds since the samples of $\mathbf{n}_G$ are independent. Inequality (5.38) holds from the property $h(x|y) \leq h(x)$, with equality if and only if $x$ and $y$ are independent. This proves statement (i) in Lemma 5.8.

- Inequality $(c)$: We have that

$$\begin{aligned}
I(\mathbf{v}_G \to \mathbf{w}_G) &= \bar{h}(\mathbf{w}_G) - h(\mathrm{w}_G(k)|\mathrm{w}_G{}_1^{k-1}, \mathrm{v}_G{}_1^{k}) \\
&= \bar{h}(\mathbf{w}_G) - h(\mathrm{w}_G(k)|\mathrm{w}_G{}_1^{k-1}, \widetilde{\mathrm{x}}_G{}_1^{k}) && (5.39)\\
&= \bar{I}(\widetilde{\mathbf{x}}_G \to \mathbf{w}_G) \\
&= \bar{I}(\widetilde{\mathbf{x}}_G; \mathbf{w}_G) && (5.40)
\end{aligned}$$

  Equality (5.39) holds since, if $\mathrm{w}_G{}_1^{k-1}$ is known, then $\widetilde{\mathrm{x}}_G{}_1^{k}$ can be obtained deterministically from $\mathrm{v}_G{}_1^{k-1}$, and vice-versa. Equality (5.40) holds from the fact that there is no feedback from $\mathbf{w}_G$ to

$\widetilde{\mathbf{x}}_G$. On the other hand, $\bar{I}(\widetilde{\mathbf{x}}_G;\mathbf{w}_G) \geq \bar{I}(\mathbf{x}_G;\mathbf{y}_G)$ with equality if and only if the null space of $\boldsymbol{B}$ is contained within the space orthogonal to the range of $\boldsymbol{A}$. This proves statement (ii) in Lemma 5.8.

- Inequality $(d)$ follows from the definition of $R_{a,b}(D)$. The conditions for equality stated in point iii) in Lemma 5.8 follow directly from Theorem 4.6.

This completes the proof. □

As in the scalar case, we can see from (5.7) that, in the scheme depicted in Fig. 5.2-(b), the quantity $\frac{1}{2N}\sum_{k=1}^{N}\log(\gamma(k)+1)$ is lower bounded by $R_{a,b}(D)$. Also, condition iii) in Lemma 5.8 can always be met, since there are no constraints on $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{F}$ other than $\boldsymbol{F}$ being zero-lower-triangular. Indeed, the combination of matrices and noise variances that solve Optimization Problem 5.4 is not unique, since achieving equality in $(d)$ of (5.34) requires to satisfy only the two matrix equations in (5.35), while there are three matrices to be chosen, two of them with complete freedom. Of course, all the combinations that solve Optimization Problem 5.4 yield $\bar{I}(\mathbf{x}_G;\mathbf{y}_G) = R_{a,b}(D)$. Moreover, it is always possible to find matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{F}$ so that all *three* conditions in Lemma 5.8 are met. As a consequence, there exists, at least, one solution to Optimization Problem 5.4 for which $\frac{1}{2N}\sum_{k=1}^{N}\log(\gamma(k)+1) = R_{a,b}(D)$.

Notice that there are no explicit requirements on the noise variances $\{\eta(k)^2\}_{k=1}^{N}$ in order to achieve equality throughout (5.34). In particular, it is not necessary that all noise variances be equal. Notice also that, in condition (i) of Lemma 5.8, *the random vector whose components need to be independent is* $\mathbf{w}$*, and not* $\mathbf{v}$. That is, *it is not required that* $\boldsymbol{A}$ *"de-correlate"* $\mathbf{x}$.

Similarly to the scalar case, the WCMSE-RDF for $\mathbf{x}_G$ also constitutes a lower bound for $\frac{1}{2N}\sum_{k=1}^{N}\log(\gamma(k)+1)$ in the not-necessarily Gaussian system shown in Fig. 5.1-(a). This is formally stated in the following lemma:

**Lemma 5.9.** *In the system depicted in Fig. 5.2-(a), the following holds:*

$$\frac{1}{2N}\sum_{k=1}^{N}\log(\gamma(k)+1) \geq R_{a,b}(D), \tag{5.41}$$

*where $R_{a,b}(D)$ is the WCMSE-RDF for the source $\mathbf{x}_G$, which is Gaussian having the same covariance matrix as $\mathbf{x}$. Equality is achieved if and only if conditions (i), (ii) and (iii) in Lemma 5.8 are met.* ▲

*Proof.* The proof is essentially the same as the proof for Lemma 5.2. □

With the above lemma, we obtain the following theorem.

**Theorem 5.10.** *Suppose there exists a triplet of matrices $[\boldsymbol{A},\boldsymbol{B},\boldsymbol{F}] \in \mathbb{F}$ that satisfy conditions (i), (ii) and (iii) in Lemma 5.8. Then, a triplet of matrices $[\boldsymbol{A}',\boldsymbol{B}',\boldsymbol{F}'] \in \mathbb{F}$ is a solution to Optimization Problem 5.3 if and only if $[\boldsymbol{A}',\boldsymbol{B}',\boldsymbol{F}']$ also satisfies conditions (i), (ii) and (iii) in Lemma 5.8.* ▲

*Proof.* If there exists a triplet of matrices $[\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{F}] \in \mathbb{F}$ that satisfies the three conditions of Lemma 5.8, then, from Lemma 5.8, $[\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{F}]$ yields $\frac{1}{2N}\sum_{k=1}^{N}\log(\gamma(k)+1) = R_{a,b}(D)$, which is the lower bound for $\frac{1}{2N}\sum_{k=1}^{N}\log(\gamma(k)+1)$ achievable with any matrices. Therefore, a triplet $[\boldsymbol{A}', \boldsymbol{B}', \boldsymbol{F}'] \in \mathbb{F}$ is a solution to Optimization Problem 5.3 only if it yields $\frac{1}{2N}\sum_{k=1}^{N}\log(\gamma(k)+1) = R_{a,b}(D)$. From Lemma 5.8, the latter holds if and only if $[\boldsymbol{A}', \boldsymbol{B}', \boldsymbol{F}']$ satisfies the three conditions of Lemma 5.8. This completes the proof. □

The next corollary follows immediately from Theorem 5.10:

**Corollary 5.11.** *Suppose there exists a triplet of matrices $[\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{F}] \in \mathbb{F}$ that satisfies conditions (i), (ii) and (iii) in Lemma 5.8. Then, every solution to Optimization Problem 5.3 is also a solution to Optimization Problem 5.4.* ▲

## Optimal Transform Coder Design

We will next apply the above results to the design of optimal transform coders [58, 120, 166]. For that purpose, it is necessary to link the quantity $\frac{1}{2N}\sum_{k=1}^{N}\log(\gamma+1)$ to the total operational bit-rate associated with the $N$ quantizers in the transform coder. More precisely, denoting the total operational bit-rate by $R_{op}$, there must exist a monotonically increasing function $L : \mathbb{R}^+ \to \mathbb{R}^+$, independent of the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{F}$, such that

$$R_{op} = L\left(\sum_{k=1}^{N}\log(\gamma+1)\right). \tag{5.42a}$$

In addition, in order to apply Lemmas 5.8, 5.9 and Theorem 5.10 to the design of optimal transform coders, we need to assume the following:

$$\mathrm{E}\left[\mathrm{v}(k)\,\mathrm{n}(k)\right] = 0, \quad \forall k \in \{1, 2, \ldots, N\} \tag{5.42ba}$$

$$\mathrm{E}\left[\mathrm{v}(k)\,\mathrm{n}(i)\right] = 0, \quad \forall k \neq i, k, i \in \{1, 2, \ldots, N\} \tag{5.42bb}$$

$$\mathrm{E}\left[\mathrm{n}(k)\,\mathrm{n}(i)\right] = \delta_{k,i}, \quad \forall k, i \in \{1, 2, \ldots, N\}. \tag{5.42bc}$$

where $\delta_{k,i}$ denotes the Kronecker delta function. The expressions in (5.42) impose requirements on the scalar quantizers for which the above results can be used. It is clear that (5.42b) can be satisfied by using dithered scalar quantizers [85, 126, 132, 134]. We will show below that dithered quantization also satisfies (5.42a).

### Suitable Scalar Quantizers

Conditions (5.42ba), (5.42bb) and (5.42bc) can be satisfied exactly by using uniform scalar quantization with dither, both subtractive and non-subtractive. For this to hold, the dither signals applied to each scalar

quantizer must be independent both mutually and from $\mathbf{v}$. Furthermore, in the case of subtractive dither, the dither has to be uniformly distributed over the quantization interval [126, 132]. In the non-subtractive case, the dither needs to either have a uniform PDF over the quantization interval or its PDF must be an $m$-fold convolution of such uniform PDFs (with $m \geq 2$), see [85, 134].

In relation to the SNR/bit-rate requirement imposed by (5.42a), we will show that by using uniform scalar quantization with dither (both subtractive ad non-subtractive), the operational rate is well approximated by the following expression:

$$R_{op} = \frac{1}{2}\sum_{k=1}^{N} \log_2(\gamma(k)+1) + \sum_{k=1}^{N} \mathscr{C}(f_{\mathrm{v}(k)}), \tag{5.3}$$

where $\mathscr{C}(\cdot)$ is a functional which depends on the type of dither technique utilized and $f_{\mathrm{v}(k)}$ is the PDF of the $k$-th component of $\mathbf{v}$. Thus, assuming that the change in the PDFs $f_{\mathrm{v}(k)}$, due to variations in $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{F}$, has a negligible net effect on $\sum_{k=1}^{N} \mathscr{C}(f_{\mathrm{v}(k)})$, the rate-SNR expression (5.3) satisfies condition (5.42a). We next discuss the validity of (5.3) for uniform scalar quantizers with dither, both subtractive and non-subtractive.

- *Entropy coded uniform scalar quantization with subtractive dither (SDUSQ)*: In this case, utilizing entropy coding conditioned on the dither, the operational rate of each scalar quantizer is:

$$r_k = \frac{1}{2}\log_2(\gamma(k)+1) + 0.254 - D(\mathrm{v}(k) \| \, \mathrm{v}_G(k)). \tag{5.4}$$

  This result follows directly from (2.58), (2.57) (on page 42) and from (4.136) (see the proof of Lemma 4.10 on page 148). Notice that (5.4) is a special case of (5.3). Thus, assuming that the effect of the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{F}$ on the divergence $D(\mathrm{v}(k) \| \, \mathrm{v}_G(k))$ can be neglected, SDUSQ satisfies all the conditions stated in (5.42). We note that the assumption that the effect of the matrices transform coder matrices on the PDFs of the subband signals is negligible is often used in the analysis and design of transform coders, both with and without feedback, see, e.g., [57, 111, 120, 165, 166].

- *Entropy coded uniform scalar quantization with triangular PDF non-subtractive dither:* The scalar entropy rate of the output of the quantizer in this case is plotted in Fig. 5.3 for Gaussian input (solid line), Laplacian input (dashed line), and uniformly distributed input (dashed dot line). It can be seen from Fig. 5.3 that for the three input PDFs considered, the entropy rates differ by not more than $0.2$ bits/sample, for all entropy rates below $5$ bits/sample (equivalently, for all $\gamma \leq 100$). Moreover, all these plots can be closely approximated by the function $\frac{1}{2}\log_2(\gamma+1)+1$, also plotted in Fig. 5.3 with dashed line and filled circle markers. The approximation error associated

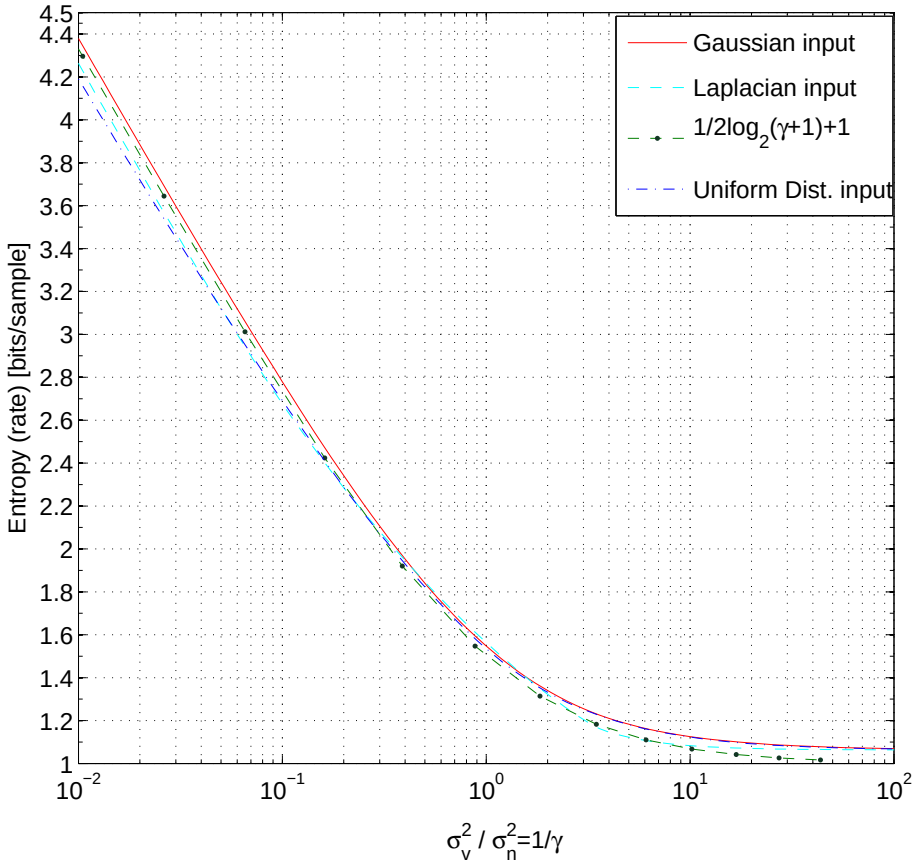



**Figure 5.3:** Output entropy rate of uniform scalar quantization with triangular non-subtractive dither.

with $\frac{1}{2}\log_2(\gamma+1)+1$ is smaller than $0.12$ bits/sample, for all input PDFs considered, and for all entropy rates below $5$ bits/sample (equivalently, for all $\gamma \leq 100$). Thus, the operational rate of each scalar quantizer can be closely approximated by:

$$r_k = \frac{1}{2}\log_2(\gamma(k)+1)+1. \tag{5.5}$$

With this approximation, uniform scalar quantization with non-subtractive triangular dither satisfies all the conditions stated in (5.42).

**Solving the Equations**

Assuming the use of dithered uniform quantization, optimal transform coders can be easily designed by using Lemma 5.9 or Theorem 5.10. From these results, it holds that, if the following three equations can be satisfied:

$$\boldsymbol{B}\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{V}^\star; \tag{5.6a}$$

$$\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{F})\,\mathrm{diag}\left\{\sigma^2_{\mathrm{n}(k)}\right\}(\boldsymbol{I}-\boldsymbol{F})^T\boldsymbol{B}^T = \boldsymbol{K}^\star_{\mathbf{u}}; \tag{5.6b}$$

$$\boldsymbol{B}\,\mathrm{diag}\left\{\sigma^2_{\mathrm{w}(k)}\right\}\boldsymbol{B}^T = \boldsymbol{K}^\star_{\mathbf{y}} \triangleq (\boldsymbol{I}-\boldsymbol{V}^\star)\boldsymbol{K}_{\mathbf{x}}(\boldsymbol{I}-\boldsymbol{V}^\star)^T + \boldsymbol{K}^\star_{\mathbf{u}} \tag{5.6c}$$

where $\boldsymbol{K}^\star_{\mathbf{y}} \triangleq (\boldsymbol{I}-\boldsymbol{V}^\star)\boldsymbol{K}_{\mathbf{x}}(\boldsymbol{I}-\boldsymbol{V}^\star)^T$, then the solution characterizes an optimal transform coder.

One possible path for solving these equation is:

1. First choose any $\boldsymbol{B}$ such that $\boldsymbol{B}^\dagger$ diagonalizes $\boldsymbol{K}^\star_{\mathbf{y}}$. This will yield $\mathrm{diag}\{\sigma^2_{\mathrm{w}(k)}\}$.
2. Then, choose $\boldsymbol{F}$ such that $(\boldsymbol{I}-\boldsymbol{F})^{-1}$ diagonalizes $\boldsymbol{B}^\dagger\boldsymbol{K}^\star_{\mathbf{u}}(\boldsymbol{B}^\dagger)^T$. This will yield $\mathrm{diag}\{\sigma^2_{\mathrm{n}(k)}\}$.
3. Finally, set $\boldsymbol{A} = \boldsymbol{B}^\dagger(\boldsymbol{I}-\boldsymbol{V}^\star)$.

Notice from (5.6a) that if $\boldsymbol{V}^\star$ is not lower triangular, then it is not possible that both $\boldsymbol{A}$ and $\boldsymbol{B}$ be lower triangular matrices. Since $\boldsymbol{V}^\star$ is symmetric for all WCMSE weights $a/b > 0$, it follows that $R_{a,b}(D)$ cannot be achieved causally unless $b \to \infty$, i.e., unless $R_{a,b}(D)$ coincides with $R^\perp(D)$.

On the other hand, using a KLT matrix followed (preceeded) by diagonal scaling matrices as $\boldsymbol{A}$ (and $\boldsymbol{B}$), constitutes a solution to (5.6) when $\boldsymbol{F} = \boldsymbol{0}$. This stems from the fact that $\boldsymbol{V}^\star$ and $\boldsymbol{K}^\star_{\mathbf{u}}$, and thus $\boldsymbol{K}^\star_{\mathbf{y}}$, are diagonalized by the same matrix. This also reveals the fact that, for random vector sources, the subband expansion inherent to transform coding can substitute the lack of feedback, effectively yielding three degrees of freedom in the design. However, if additional constraints are imposed on $\boldsymbol{A}$ or $\boldsymbol{B}$, then feedback becomes necessary for optimality, as illustrated in the following situation.

**Causal Transform Coder Design Example.** In order to design an optimal perfect reconstruction causal transform coder, one must first determine the value of the scalar parameter $\alpha$ in (4.61a), by solving (4.61a) (or (4.61b)) for the desired rate (or distortion). Then, the matrices $\boldsymbol{K}_{\mathrm{u}}^{\star}$ and $\boldsymbol{V}^{\star}$ can be obtained from (4.62a) and (4.62b), respectively. Following this, the matrix $\boldsymbol{B}$ can be chosen as:

$$\boldsymbol{B} = \boldsymbol{L}\left(\operatorname{diag}\left\{\sigma^2_{\mathrm{w}(k)}\right\}\right)^{1/2}, \tag{5.7}$$

where $\boldsymbol{L}$ is the a lower triangular matrix such that

$$\boldsymbol{K}_{\mathbf{y}}^{\star} = \boldsymbol{L}\boldsymbol{L}^T \tag{5.8}$$

is a Cholesky decomposition of $\boldsymbol{K}_{\mathbf{y}}^{\star} = (\boldsymbol{I} - \boldsymbol{V}^{\star})\boldsymbol{K}_{\mathbf{x}}(\boldsymbol{I} - \boldsymbol{V}^{\star})^T$ [159]. Then choose $\boldsymbol{A} = \boldsymbol{B}^{-1}$, which is lower triangular. Finally, choose $(\boldsymbol{I} - \boldsymbol{F})$ to diagonalize $\boldsymbol{B}^{-1}\boldsymbol{K}_{\mathbf{u}}^{\star}\boldsymbol{B}^{-T}$. Such a choice of $(\boldsymbol{I} - \boldsymbol{F})$, which is constrained to be a unit-lower triangular matrix, always exists, and can be found by using, e.g., [41, Lemma 1]. Since $\boldsymbol{A}$ and $\boldsymbol{B}$ are lower triangular, the resulting transform coder is causal. (Indeed, it is zero-delay.) We note that $\boldsymbol{K}_{\mathbf{u}}^{\star}$ and $\boldsymbol{K}_{\mathbf{y}}^{\star}$ cannot be diagonalized by the same triangular matrix unless $\boldsymbol{K}_{\mathbf{u}}^{\star} = \beta\boldsymbol{K}_{\mathbf{y}}^{\star}$, for some scalar $\beta > 0$. This implies that, *in order to obtain an optimal causal transform coder, the use of a feedback matrix (and hence of all three degrees of freedom) is necessary*. Notice also that, in an optimal causal transform coder, *the components of* $\mathbf{v}$ *are not uncorrelated*.

For Gaussian sources, and for any source-uncorrelated reconstruction MSE value $D > 0$, the operational rate-distortion performance of the causal transform coder obtained from the above equations will exceed $R_{1,\infty}(D)$ by $\sum_{k=1}^{N}\mathscr{C}(f_{\mathrm{v}(k)})$, see (5.3). In particular, if entropy coded SDUSQ is used, the operational bit-rate will exceed $R_{a,b}(D)$ by not more than $0.254$ bits/sample (see (2.60) on page 42). Moreover, if the variations of $\sum_{k=1}^{N}\mathscr{C}(f_{\mathrm{v}(k)})$ produced by different choices of matrices is negligible, then the obtained causal transform coder will be optimal within the family of all transform coders using scalar quantizers with the same rate-SNR function. This means that, using entropy coded dithered quantizers and with the matrices obtained from (5.6), causal transform coding is as rate-distortion efficient as non-causal PR transform coding, *at all rates*.

## 5.4 Conditions for Vector Processes

In this section we extend the results derived in sections 5.2 and 5.3 to the cases in which the source is an $N$-dimensional vector process. For this purpose, consider the system shown in Fig. 5.4-(a). In this figure, $\{\mathbf{x}(k)\}$ and $\{\mathbf{n}(k)\}$ are zero mean, jointly w.s.s. random vector processes, and $\boldsymbol{A}(z)$, $\boldsymbol{B}(z)$ and $\boldsymbol{F}(z)$ are $N \times N$ transfer functions matrices. The $N$ parallel processes $\{\mathrm{n}_i(k)\}$, $i = 1, 2, \ldots, N$, which

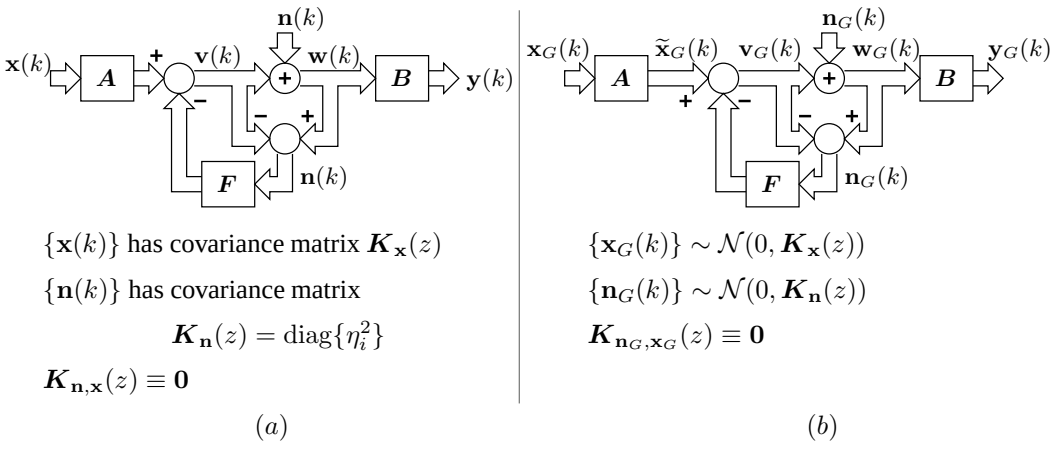



**Figure 5.4:** a) Linear model of an encoder for vector processes. b) Forward vector test channel.

comprise $\{\mathbf{n}(k)\}$, are white, mutually uncorrelated, and $\{\mathrm{n}_i(k)\}$ is uncorrelated with $\{\mathrm{x}_j(k)\}$, for all $i \neq j \in \{1, 2, \ldots, N\}$. More precisely, we have

$$\boldsymbol{K}_{\mathbf{n}}(z) = \mathrm{diag}\left\{\eta_i^2\right\}, \quad \forall i = 1, 2, \ldots, N \tag{5.9}$$

$$\boldsymbol{K}_{\mathbf{n},\mathbf{x}}(z) \equiv \mathbf{0}, \tag{5.10}$$

where

$$\eta_i^2 \triangleq \mathrm{E}\left[\mathrm{n}_i(k)^2\right], \quad i = 1, 2, \ldots, N, \tag{5.11}$$

is the variance of the $i$-th process in $\{\mathbf{n}(k)\}$.

The system in Fig. 5.4-(a) could be regarded as an analysis model for an ED pair for vector processes. If the processes $\{\mathrm{x}_i(k)\}$ are the result of a polyphase decomposition of a scalar random process, this model can represent a filter bank with feedback [28, 61, 115, 117, 167]. In this case, and using the Linear Model, $\{\mathbf{n}(k)\}$ would represent the error processes introduced by $N$ parallel scalar quantizers. As in the transform coder case, in order to avoid algebraic loops, each element in $\boldsymbol{F}(z)$ must be a causal transfer function, and $\boldsymbol{F}(z)$ needs to be lower triangular with zeros along its main diagonal (i.e., $\boldsymbol{F}(z)$ needs to be *strictly causal*).

The *vector of signal-to-noise ratios* in this case is defined as

$$\boldsymbol{\gamma} \triangleq [\gamma_1, \gamma_2, \ldots, \gamma_N]^T, \tag{5.12}$$

where

$$\gamma_i \triangleq \frac{\sigma_{\mathrm{v}_i}^2}{\eta_i^2}, \quad i = 1, 2, \ldots, N \tag{5.13}$$

denotes the scalar SNR in the $i$-th channel, and were $\sigma^2_{\mathrm{v}_i}$ is the variance of the $i$-th process of $\{\mathbf{v}(k)\}$.

Under the above conditions, we can define the following optimization problem:

**Optimization Problem 5.5.** *For the system depicted in Fig. 5.4-(a),*

$$\textit{Minimize: } \frac{1}{2N}\sum_{i=1}^{N}\log_2(\gamma_i+1) \tag{5.14}$$

$$\textit{Subject to: } D_{a,b}(\{\mathbf{x}(k)\},\{\mathbf{y}(k)\})\leq D, \tag{5.15}$$

*over all w.s.s. processes $\{\mathbf{n}(k)\}$ and over all transfer functions $\boldsymbol{A}(z)$, $\boldsymbol{B}(z)$ and $\boldsymbol{F}(z)$ such that $\boldsymbol{F}(z)$ is strictly causal and $[\boldsymbol{A}(z),\boldsymbol{B}(z),\boldsymbol{F}(z)]\in\mathbb{F}$, where $\mathbb{F}$ is a constraint set of matrix transfer function triplets.* ▲

The second system, depicted in Fig. 5.4-(b), has the same transfer functions as the system of Fig. 5.4-(a), but differs from the latter in that its source and noise are the Gaussian random vector processes $\{\mathbf{x}_G(k)\}$ and $\{\mathbf{n}_G(k)\}$, having the same covariance matrices as $\{\mathbf{x}(k)\}$ and $\{\mathbf{n}(k)\}$, respectively. Since both $\{\mathbf{x}_G(k)\}$ and $\{\mathbf{n}_G(k)\}$ are Gaussian, the no-correlation condition $\boldsymbol{K}_{\mathbf{n}_G,\mathbf{x}G}(z)\equiv\mathbf{0}$ implies that $\{\mathbf{x}_G(k)\}$ and $\{\mathbf{n}_G(k)\}$ are independent. Similarly, the fact that $\boldsymbol{K}_{\mathbf{n}_G}(z)$ is diagonal implies that the scalar processes in $\{\mathbf{n}_G(k)\}$ are mutually independent. The system can then be seen as a forward vector channel realization of the WCMSE-RDF for Gaussian vector process sources, characterized in Section 4.6. From Definition 4.5, the transfer function matrices $\boldsymbol{A}(z)$, $\boldsymbol{B}(z)$ and $\boldsymbol{F}(z)$ that yield a realization of $R_{a,b}(D)$ for $\{\mathbf{x}_G(k)\}$ must necessarily solve the following optimization problem:

**Optimization Problem 5.6.** *For the system depicted in Fig. 5.4-(b),*

$$\textit{Minimize: } \bar{I}(\{\mathbf{x}_G(k)\};\{\mathbf{y}_G(k)\}) \tag{5.16}$$

$$\textit{Subject to: } D_{a,b}(\{\mathbf{x}_G(k)\},\{\mathbf{y}_G(k)\})\leq D, \tag{5.17}$$

*over all stationary Gaussian vector processes $\{\mathbf{n}(k)\}$ and over all transfer function matrices $\boldsymbol{A}(z)$, $\boldsymbol{B}(z)$ and $\boldsymbol{F}(z)$ such that $\boldsymbol{F}(z)$ is strictly causal.* ▲

Clearly, the vectors of SNRs in both systems in Fig. 5.4 are the same, since, in both systems, corresponding signals have the same second moments.

Before extending lemmas 5.1 and 5.8 to vector processes, it is convenient to note that, for the system of Fig 5.4-(b), the directed mutual information rate from $\{\mathbf{v}(k)\}$ to $\{\mathbf{w}(k)\}$ takes the following form:

$$\bar{I}(\{\mathbf{v}(k)\}\to\{\mathbf{w}(k)\})=\bar{I}(\mathbf{v}^k\to\mathbf{w}(k)|\mathbf{w}^{k-1})=\sum_{i=1}^{N}I\left(\mathbf{v}^{k-1},\mathbf{v}(k)_1^i\,;\,\mathrm{w}_i(k)|\mathbf{w}^{k-1},\mathbf{w}(k)_1^{i-1}\right) \tag{5.18}$$

Using this fact, we can state the following result:

**Lemma 5.12.** *In the system depicted in Fig. 5.4-(b), the following holds:*

$$\frac{1}{2N}\sum_{k=1}^{N}\log(\gamma_i+1) \overset{(a)}{=} \frac{1}{N}\sum_{i=1}^{N} I(\mathrm{v}_{Gi}(k);\mathrm{w}_{Gi}(k)) \overset{(b)}{\geq} \bar{I}(\{\mathbf{v}_G(k)\}\to\{\mathbf{w}_G(k)\}) \tag{5.19}$$

$$\overset{(c)}{\geq} \bar{I}(\{\mathbf{x}_G(k)\};\{\mathbf{y}_G(k)\}) \overset{(d)}{\geq} R_{a,b}(D). \tag{5.20}$$

*In addition,*

*i) Equality is achieved in* $(b)$ *if and only if* $\boldsymbol{K}_{\mathbf{w}_G}(\mathrm{e}^{j\omega})$ *is a constant diagonal matrix.*

*ii) Equality is achieved in* $(c)$ *if and only if* $\mathcal{N}_{\boldsymbol{B}} \subseteq \mathcal{R}_{\boldsymbol{A}}^{\perp}$.

*iii) Equality is achieved in* $(d)$ *iff*

$$\boldsymbol{B}(\mathrm{e}^{j\omega})(\boldsymbol{I}-\boldsymbol{F}(\mathrm{e}^{j\omega}))\boldsymbol{K}_{\mathbf{n}_G}(\mathrm{e}^{j\omega})(\boldsymbol{I}-\boldsymbol{F}(\mathrm{e}^{j\omega}))^H\boldsymbol{B}(\mathrm{e}^{j\omega})^H = \boldsymbol{K}_{\mathbf{u}}^{\star}(\mathrm{e}^{j\omega}) \tag{5.21a}$$

$$\boldsymbol{B}(\mathrm{e}^{j\omega})\boldsymbol{A}(\mathrm{e}^{j\omega}) = \boldsymbol{I}-\boldsymbol{V}^{\star}(\mathrm{e}^{j\omega}), \tag{5.21b}$$

*where* $\boldsymbol{K}_{\mathbf{u}}^{\star}(\mathrm{e}^{j\omega})$ *and* $\boldsymbol{V}^{\star}(\mathrm{e}^{j\omega})$ *are as defined in* (4.121).

▲

*Proof.* We proceed by parts.

- Equality $(a)$ follows from the fact that $\{\mathbf{x}_G(k)\}$ and $\{\mathbf{n}_G(k)\}$ are independent Gaussian stationary random vector processes, together with the fact that $\boldsymbol{F}(z)$ is strictly causal.

- Inequality $(b)$: We have that

$$\begin{aligned}
\sum_{i=1}^{N} I(\mathrm{v}_{Gi}(k);\mathrm{w}_{Gi}(k)) &= \sum_{i=1}^{N} h(\mathrm{w}_{Gi}(k)) - h(\mathrm{w}_{Gi}(k)|\,\mathrm{v}_{Gi}(k)) \\
&= \sum_{i=1}^{N} h(\mathrm{w}_{Gi}(k)) - h(\mathrm{n}_{Gi}(k)|\,\mathrm{v}_{Gi}(k)) \\
&\overset{(e)}{=} \sum_{i=1}^{N} h(\mathrm{w}_{Gi}(k)) - h(\mathrm{n}_{Gi}(k)) \\
&\overset{(f)}{=} \sum_{i=1}^{N} h(\mathrm{w}_{Gi}(k)) - h\left(\mathrm{n}_{Gi}(k)|\mathbf{w}_G^{k-1},\mathbf{w}_G(k)_1^{i-1},\mathbf{v}_G^{k-1},\mathbf{v}_G(k)_1^{i}\right) \\
&= \sum_{i=1}^{N} h(\mathrm{w}_{Gi}(k)) - h\left(\mathrm{w}_{Gi}(k)|\mathbf{w}_G^{k-1},\mathbf{w}_G(k)_1^{i-1},\mathbf{v}_G^{k-1},\mathbf{v}_G(k)_1^{i}\right) \\
&\overset{(g)}{\geq} \sum_{i=1}^{N} h(\mathrm{w}_{Gi}(k)|\mathbf{w}^{k-1},\mathbf{w}(k)_1^{i-1}) - h\left(\mathrm{w}_{Gi}(k)|\mathbf{w}_G^{k-1},\mathbf{w}_G(k)_1^{i-1},\mathbf{v}_G^{k-1},\mathbf{v}_G(k)_1^{i}\right) \\
&= \sum_{i=1}^{N} I\left(\mathbf{v}_G^{k-1},\mathbf{v}_G(k)_1^{i}\,;\,\mathrm{w}_{Gi}(k)|\mathbf{w}_G^{k-1},\mathbf{w}_G(k)_1^{i-1},\mathbf{v}_G^{k-1},\mathbf{v}_G(k)_1^{i}\right) \\
&\overset{(h)}{=} \bar{I}(\{\mathbf{v}_G(k)\}\rightarrow\{\mathbf{w}_G(k)\})
\end{aligned}$$

  In the above, equality $(e)$ follows from the fact that $\{\mathbf{n}_G(k)\}$ and $\{\mathbf{x}_G(k)\}$ are independent and from the fact that $\boldsymbol{F}(z)$ is strictly causal. As a consequence, $\mathrm{n}_{Gi}(k)$ is independent of $\mathrm{v}_{Gi}(k)$, for all $i\in\{1,\ldots,N\}$. Similarly, equality $(f)$ holds since $\mathbf{n}_{Gi}(k)$ is independent of $\mathbf{w}_G^{k-1}$, $\mathbf{w}_G(k)_1^{i-1}$, $\mathbf{v}_G^{k-1}$ and $\mathbf{v}_G(k)_1^{i}$, $\forall i\in\{1,2,\ldots,N\}$, $\forall k\in\mathbb{Z}$. Inequality $(g)$ holds from the property $h(x|y)\leq h(x)$, with equality if and only if $\boldsymbol{K}_{\mathbf{y}}(\mathrm{e}^{j\omega})$ is a constant diagonal matrix. Equality $(h)$ follows directly from (5.18). This proves statement (i) in Lemma 5.12.

- Inequality $(c)$: We have that

$$\begin{aligned}
&\bar{I}(\{\mathbf{v}_G(k)\}\rightarrow\{\mathbf{w}_G(k)\}) \\
&\qquad = h(\mathrm{w}_{Gi}(k)|\mathbf{w}^{k-1},\mathbf{w}(k)_1^{i-1}) - h\left(\mathrm{w}_{Gi}(k)|\mathbf{w}_G^{k-1},\mathbf{w}_G(k)_1^{i-1},\mathbf{v}_G^{k-1},\mathbf{v}_G(k)_1^{i}\right) \\
&\qquad \overset{(i)}{=} h(\mathrm{w}_{Gi}(k)|\mathbf{w}^{k-1},\mathbf{w}(k)_1^{i-1}) - h\left(\mathrm{w}_{Gi}(k)|\mathbf{w}_G^{k-1},\mathbf{w}_G(k)_1^{i-1},\widetilde{\mathbf{x}}_G^{k-1},\widetilde{\mathbf{x}}_G(k)_1^{i}\right) \\
&\qquad = \bar{I}(\{\widetilde{\mathbf{x}}_G(k)\}\rightarrow\{\mathbf{w}_G(k)\}) \\
&\qquad \overset{(j)}{=} \bar{I}(\{\widetilde{\mathbf{x}}_G(k)\};\{\mathbf{w}_G(k)\})
\end{aligned}$$

  where equality $(i)$ follows from the fact that, if $\mathbf{w}_G^{k-1}$ and $\mathbf{w}(k)_1^{i-1}$ are known, then $\widetilde{\mathbf{x}}_G^{k-1}$ and $\widetilde{\mathbf{x}}_G(k)_1^{i}$ can be obtained deterministically from ${\mathbf{v}_G}^{k-1}$ and $\mathbf{v}_G(k)_1^{i}$, and vice-versa. Equality $(j)$ follows from the fact that there exists no feedback from $\{\mathbf{w}_G(k)\}$ to $\{\widetilde{\mathbf{x}}_G(k)\}$. On the other hand,

$\bar{I}(\{\widetilde{\mathbf{x}}_G(k)\};\{\mathbf{w}_G(k)\}) \geq \bar{I}(\{\mathbf{x}_G(k)\};\{\mathbf{y}_G(k)\})$ (data processing inequality), with equality if and only if the null space of $\boldsymbol{B}$ is contained in the space orthogonal to the range of $\boldsymbol{A}$. This proves statement (ii) in Lemma 5.12.

- Inequality $(d)$ follows from the definition of $R_{a,b}(D)$. The conditions for equality stated in point iii) in Lemma 5.12 follow directly from Theorem 4.8.

This completes the proof. □

Since in Optimization Problem 5.6 the transfer function matrices $\boldsymbol{A}(z)$, $\boldsymbol{B}(z)$ and $\boldsymbol{F}(z)$ are not subject to any constraint, aside from $\boldsymbol{F}(z)$ being strictly causal, it follows that condition iii) in Lemma 5.12 can always be satisfied. Moreover, there exist infinite combinations of transfer function matrices and noise variances that solve Optimization Problem 5.6. Of course, all these combinations yield $\bar{I}(\{\mathbf{x}_G(k)\};\{\mathbf{y}_G(k)\}) = R_{a,b}(D)$.

Of greater practical importance, we note that it is always possible to find transfer function matrices $\boldsymbol{A}(z)$, $\boldsymbol{B}(z)$ and $\boldsymbol{F}(z)$ so as to satisfy all *three* conditions in Lemma 5.12. This means that there exists at least one solution to Optimization Problem 5.6 for which $\frac{1}{2N}\sum_{i=1}^{N}\log(\gamma_i+1) = R_{a,b}(D)$.

Notice that there are no explicit requirements on the noise variances $\{\eta_i^2\}_{i=1}^N$ in order to achieve equality throughout (5.19). In particular, it is not necessary that all noise variances be equal. Notice also that, in condition (i) of Lemma 5.12, *it is* $\{\mathbf{w}(k)\}$, *and not* $\{\mathbf{v}(k)\}$, the random vector process whose components need to be independent. More precisely, when the strictly causal feedback matrix transfer function $\boldsymbol{F}(z)$ can be chosen freely, then *it is not required that* $\boldsymbol{A}(z)$ *de-correlates the processes within* $\{\mathbf{x}(k)\}$.

The result stated by Lemma 5.12 for the Gaussian system in Fig. 5.4-(b) has an important implication in the not-necessarily-Gaussian system of Fig. 5.4-(a), as stated in the following lemma:

**Lemma 5.13.** *In the system depicted in Fig. 5.4-(a), the following holds:*

$$\frac{1}{2N}\sum_{i=1}^{N}\log(\gamma_i+1) \geq R_{a,b}(D), \tag{5.22}$$

*where $R_{a,b}(D)$ is the WCMSE-RDF for a source $\{\mathbf{x}_G(k)\}$, which is Gaussian, stationary, and has the same covariance matrix as $\{\mathbf{x}(k)\}$. Equality is achieved if and only if conditions (i), (ii) and (iii) in Lemma 5.12 are met.* ▲

*Proof.* The proof is essentially the same as the proof for Lemma 5.2. □

The previous lemma allows one to state the following result:

**Theorem 5.14.** *Suppose there exists a triplet of transfer function matrices* $[\boldsymbol{A}(z), \boldsymbol{B}(z), \boldsymbol{F}(z)] \in \mathbb{F}$ *that satisfy conditions (i), (ii) and (iii) in Lemma 5.12. Then, a triplet of matrix transfer function* $[\boldsymbol{A}'(z), \boldsymbol{B}'(z), \boldsymbol{F}'(z)] \in \mathbb{F}$ *is a solution to Optimization Problem 5.5 if and only if* $[\boldsymbol{A}'(z), \boldsymbol{B}'(z), \boldsymbol{F}'(z)]$ *also satisfies conditions (i), (ii) and (iii) in Lemma 5.12.* ▲

*Proof.* If there exists a triplet of transfer function matrices, say $[\boldsymbol{A}(z), \boldsymbol{B}(z), \boldsymbol{F}(z)] \in \mathbb{F}$, that satisfies the three conditions of Lemma 5.12, then, from Lemma 5.12, $[\boldsymbol{A}(z), \boldsymbol{B}(z), \boldsymbol{F}(z)]$ yields $\frac{1}{2N}\sum_{i=1}^{N}\log(\gamma_i + 1) = R_{a,b}(D)$, which is the lower bound for $\frac{1}{2N}\sum_{i=1}^{N}\log(\gamma_i + 1)$ achievable by any transfer function matrices. Therefore, a triplet $[\boldsymbol{A}'(z), \boldsymbol{B}'(z), \boldsymbol{F}'(z)] \in \mathbb{F}$ is solution to Optimization Problem 5.5 only if it yields $\frac{1}{2N}\sum_{i=1}^{N}\log(\gamma_i + 1) = R_{a,b}(D)$. From Lemma 5.12, the latter holds if and only if $[\boldsymbol{A}'(z), \boldsymbol{B}'(z), \boldsymbol{F}'(z)]$ satisfies the three conditions of Lemma 5.12. This completes the proof. □

The next corollary follows immediately from Theorem 5.14.

**Corollary 5.15.** *Suppose there exists a triplet of transfer function matrices* $[\boldsymbol{A}(z), \boldsymbol{B}(z), \boldsymbol{F}(z)] \in \mathbb{F}$ *that satisfies conditions (i), (ii) and (iii) in Lemma 5.12. Then, every solution to Optimization Problem 5.5 is also a solution to Optimization Problem 5.6.* ▲

## Optimal Filter Bank Design

If the vector process $\{\mathbf{x}(k)\}$ originates from the polyphase transformation of a scalar process, then $\boldsymbol{A}(z)$ and $\boldsymbol{B}(z)$ would constitute the analysis and synthesis polyphase matrices of a filter bank [59, 60, 168]. If independent scalar quantizers are utilized in the subbands, with either subtractive dither or triangular non-subtractive dither, together with memoryless entropy coding, then the operational rate can be assumed to be a monotonically increasing function of $\frac{1}{2N}\sum_{i=1}^{N}\log_2(\gamma_i+1)$ (see the results discussed in Section 5.3). In these cases, the results obtained in this section lead directly to the optimal choice for $\boldsymbol{A}(z)$, $\boldsymbol{B}(z)$ and $\boldsymbol{F}(z)$, as discussed next.

It follows from Theorem 5.14 that, if the following three equations can be satisfied,

$$\boldsymbol{I} - \boldsymbol{V}^{\star}(\mathrm{e}^{j\omega}) = \boldsymbol{B}(\mathrm{e}^{j\omega})\boldsymbol{A}(\mathrm{e}^{j\omega}); \tag{5.23a}$$

$$\boldsymbol{K}^{\star}_{\mathbf{u}}(\mathrm{e}^{j\omega}) = \boldsymbol{B}(\mathrm{e}^{j\omega})(\boldsymbol{I} - \boldsymbol{F}(\mathrm{e}^{j\omega}))\,\mathrm{diag}\left\{\sigma^2_{\mathrm{n}_i}\right\}(\boldsymbol{I} - \boldsymbol{F}(\mathrm{e}^{j\omega}))^T\boldsymbol{B}(\mathrm{e}^{j\omega})^H; \tag{5.23b}$$

$$\boldsymbol{B}(\mathrm{e}^{j\omega})\,\mathrm{diag}\left\{\sigma^2_{\mathrm{w}_i}\right\}\boldsymbol{B}(\mathrm{e}^{j\omega})^H = \boldsymbol{K}^{\star}_{\mathbf{y}}(\mathrm{e}^{j\omega}) \triangleq (\boldsymbol{I} - \boldsymbol{V}^{\star}(\mathrm{e}^{j\omega}))\boldsymbol{K}_{\mathbf{x}}(\mathrm{e}^{j\omega})(\boldsymbol{I} - \boldsymbol{V}^{\star}(\mathrm{e}^{j\omega}))^H + \boldsymbol{K}^{\star}_{\mathbf{u}}(\mathrm{e}^{j\omega}) \tag{5.23c}$$

then the solution will characterize an optimal filter bank. For any given target rate or distortion, which will yield $\boldsymbol{K}^{\star}_{\mathbf{u}}(\mathrm{e}^{j\omega})$ and $\boldsymbol{V}^{\star}(\mathrm{e}^{j\omega})$ via (4.121), a possible path for solving (5.23) is the following:

1. First choose any $\boldsymbol{B}(\mathrm{e}^{j\omega})$ such that $\boldsymbol{B}(\mathrm{e}^{j\omega})^{\dagger}\boldsymbol{K}^{\star}_{\mathbf{y}}(\mathrm{e}^{j\omega})[\boldsymbol{B}(\mathrm{e}^{j\omega})^{\dagger}]^{H}$ yields a constant diagonal matrix. This matrix will be $\mathrm{diag}\{\sigma^2_{\mathrm{w}_i}\}$.

2. Then, choose $\boldsymbol{F}(\mathrm{e}^{j\omega})$ such that

$$(\boldsymbol{I}-\boldsymbol{F}(\mathrm{e}^{j\omega}))^{-1}\boldsymbol{B}(\mathrm{e}^{j\omega})^{\dagger}\boldsymbol{K}^{\star}_{\mathbf{u}}(\mathrm{e}^{j\omega})(\boldsymbol{B}(\mathrm{e}^{j\omega})^{\dagger})^{H}(\boldsymbol{I}-\boldsymbol{F}(\mathrm{e}^{j\omega}))^{-H}$$

   gives a constant diagonal matrix. This will be $\mathrm{diag}\{\sigma^2_{\mathrm{n}_i}\}$.

3. Finally, set $\boldsymbol{A}(\mathrm{e}^{j\omega})=\boldsymbol{B}(\mathrm{e}^{j\omega})^{\dagger}(\boldsymbol{I}-\boldsymbol{V}^{\star}(\mathrm{e}^{j\omega}))$.

Notice that, in general, if no feedback is used (i.e., if $\boldsymbol{F}(z)\equiv\mathbf{0}$), then optimal performance cannot be attained. To see this, notice that, if $\boldsymbol{B}(\mathrm{e}^{j\omega})$ is such that $\boldsymbol{B}(\mathrm{e}^{j\omega})^{\dagger}\boldsymbol{K}^{\star}_{\mathbf{y}}(\mathrm{e}^{j\omega})[\boldsymbol{B}(\mathrm{e}^{j\omega})^{\dagger}]^{H}$ is a constant matrix, then $\boldsymbol{B}(\mathrm{e}^{j\omega})^{\dagger}\boldsymbol{K}^{\star}_{\mathbf{u}}(\mathrm{e}^{j\omega})[\boldsymbol{B}(\mathrm{e}^{j\omega})^{\dagger}]^{H}$ will not be a constant matrix, unless $\boldsymbol{K}_{\mathbf{x}}(\mathrm{e}^{j\omega})$ is constant, see (4.121c). Thus, without feedback, all three conditions in Lemma 5.12 cannot be simultaneously met. In view of this observation, Theorem 5.14 implies that, under the Linear Model, and with the operational bit-rate depending on the SNRs of the quantizers in each subband as in (5.42a), not using feedback is rate-distortion suboptimal. More generally, it is easy to see that, except for special cases, not being able to exploit any of the three degrees of freedom discussed in Section 1.1.3, herein embodied in $\boldsymbol{A}(z)$, $\boldsymbol{B}(z)$ and $\boldsymbol{F}(z)$, entails a rate-distortion penalty.

On the other hand, when feedback is used, it is necessary for optimality that *the signals entering each scalar quantizer are mutually correlated*. This stands in stark contrast with the case of subband coding without feedback, where it has been shown that un-correlation between subband signals before quantization is a necessary condition for optimality, see, e.g., [26, 86, 113]. Notice also that the optimal filter bank, obtained by solving (5.23), does not necessarily satisfy the majorization property. This property consists of having the spectral densities of the scalar processes in $\{\mathbf{v}(k)\}$, say $S_{\mathrm{v}_i}(\mathrm{e}^{j\omega})$, to satisfy

$$S_{\mathrm{v}_{p(m)}}(\mathrm{e}^{j\omega})\geq S_{\mathrm{v}_{p(n)}}(\mathrm{e}^{j\omega}),\qquad \forall\omega\in[-\pi,\pi],\forall N\geq m\geq n\geq 1, \tag{5.24}$$

for some permutation $p(\cdot)$. Majorization has been shown to be a necessary condition for optimality in subband coders without feedback, see, e.g., [26, 86, 113]. The fact that it is not necessary when unconstrained feedback is used can be explained by noting that re-sorting spectral components of the source as in (5.24) yields subband signals having a flatter PSD. This is beneficial, since the rate-distortion efficiency of scalar quantization (possibly with memoryless entropy coding) is increased as the spectrum of the signal being quantized becomes more flatter [55]. However, when optimal feedback is used together with scalar quantization, the resulting performance is not dependent on the spectral density of the source, see Remark 4.4 on page 135.

## 5.5 Summary

In this chapter, sufficient conditions have been derived under which knowledge of a realization of the WCMSE-rate-distortion function can be directly used to obtain the optimal linear processing elements around scalar quantizers. The conditions established for w.s.s. scalar processes, in Section 5.2, lead directly to the optimal filters characterized for feedback quantizers characterized in Section 3.9. In the vector case, which was treaded in Section 5.3, these conditions yield the matrices for rate-distortion optimal transform coders utilizing dithered scalar quantizers. In Section 5.4, we have briefly illustrated how the conditions derived for the case of w.s.s. random vector process sources directly yield the optimal transfer function matrices in a filter bank. It has also been shown that, under a linear model of quantization errors, and except for particular cases, feedback is necessary to attain rate-distortion optimality in filter bank encoder-decoder pairs. Under these assumptions, it was also shown that for optimality (which requires the use of feedback), the majorization property is not necessary. In particular, it is not necessary in perfect reconstruction (PR) FBs. It was also shown in Section 5.4 that, under the Linear Model, filter banks in which the subband signals are mutually uncorrelated (prior to quantization) are not optimal. These two observations stand in stark contrast with what is obtained for filter banks that do not use feedback, see, e.g., [86, 113].

# Chapter 6

# Bounds to the Causal Rate-Distortion Function for Gaussian Processes

*Never do today what you can put off till tomorrow.*
*Delay may give clearer light as to what is best to be done.*
*Aaron Burr, former vice-president of the United States of America.*

*If you make delay even ambrosia turns into poison.*
*Telugu proverb.*

## 6.1 Introduction

The operation of an encoder-decoder pair consists of encoding $\{\mathrm{x}(k)\}$ into a binary sequence, which is then decoded to generate the reconstruction $\{\mathrm{y}(k)\}$. The end-to-end effect of any ED pair can be described by a series of *reproduction functions* $\{f_k\}_{k=1}^{\infty}$, such that, for every $k \in \mathbb{Z}^+$,

$$\mathrm{y}_1^k = f_k(\mathrm{x}_1^\infty). \tag{6.1}$$

As already outlined in Section 1.2.3, an encoder-decoder pair is deemed causal if the reconstruction of the current sample in the decoder is a function *only* of the current and past samples of the source, see [125]. To be more precise, we adopt the following definition from [125] [1]:

[1] The analysis in [125] considers two-sided source processes $\{\mathrm{x}(k)\}_{k=-\infty}^{\infty}$, and the reconstruction of the source samples $\{\mathrm{x}(k)\}_1^\infty$ only. Here we restrict to one-sided source processes to facilitate the connection with the definition of entropy rate, see Definition 2.13 in Section 2.3.1.

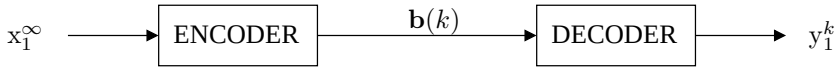



**Figure 6.1:** Representation of an ED pair at the instant the output sequence $\mathrm{y}_1^k$ is generated by the decoder.

**Definition 6.1** (Causal Source Coder)**.** *An ED is said to be causal if and only if its reproduction functions are such that*

$$f_k(\mathrm{x}_1^\infty) = f_k(\tilde{\mathrm{x}}_1^\infty), \quad \textit{whenever } \mathrm{x}_1^k = \tilde{\mathrm{x}}_1^k, \quad \forall k \in \mathbb{Z}^+ \tag{6.2}$$

▲

Thus, the fact that a given ED pair is causal can be made more explicit by re-writing (6.1) as

$$\mathrm{y}_1^k = f_k(\mathrm{x}_1^k), \quad \forall k \in \mathbb{Z}^+. \tag{6.3}$$

It also follows from Definition 6.1 and Definition 2.18 (on page 39) that an ED pair is causal if and only if the following Markov chain holds:

$$\mathrm{x}_1^\infty \to \mathrm{x}_1^k \to \mathrm{y}_1^k, \quad \forall k \in \mathbb{Z}. \tag{6.4}$$

This can be easily seen by noticing that the Markov chain is equivalent to the conditional independence situation $f_{\mathrm{y}_1^k, \mathrm{x}_1^\infty \,|\, \mathrm{x}_1^k}(y_1^k, x_1^\infty | x_1^k) = f_{\mathrm{y}_1^k \,|\, \mathrm{x}_1^k}(y_1^k|x_1^k) f_{\mathrm{x}_1^\infty \,|\, \mathrm{x}_1^k}(x_1^\infty|x_1^k)$, i.e., upon knowledge of $\mathrm{x}_1^k$, it holds that $\mathrm{y}_1^k$ is independent of $\mathrm{x}_1^\infty$ and, in particular, independent of $\mathrm{x}_{k+1}^\infty$. Having that $\mathrm{y}_1^k$ is independent of $\mathrm{x}_{k+1}^\infty$, upon knowing $\mathrm{x}_1^k$, is a necessary and sufficient condition for (6.2) to hold.

We define $\mathrm{L}_k(\mathrm{x}_1^\infty)$ to be the total number of bits that the decoder has received when it generates the output subsequence $\mathrm{y}_1^k$. Let $\mathbf{b}(k) \in \{0,1\}^{\mathrm{L}(k)}$ be the random binary sequence that contains the bits that the decoder has received when $\mathrm{y}_1^k$ is generated. Notice that $\mathrm{L}_k$ is, in general, a function of all source samples, since the binary coding may be non-causal, i.e., $\mathrm{y}_1^k$ may be generated only after the decoder has received enough bits to reproduce $\mathrm{y}_1^m$, where $m \geq k$. This is illustrated in Fig. 6.1. We highlight the fact that even though $\mathbf{b}(k)$ may contain bits which depend on samples $\mathrm{x}_{k+1}^m$ with $m > k$, the sequences $\mathrm{x}_1^\infty$ and $\mathrm{y}_1^k$ may still satisfy (6.4), i.e., the ED pair can still be causal. Notice also that $\mathrm{L}_k(\mathrm{x}_1^\infty)$ is a random variable, which depends on $\mathrm{x}_1^\infty$, the functions $\{f_k\}$ and on the manner in which the ED encodes the source into the binary sequence sent to the decoder.

For further analysis, we define the *average operational rate* of an ED pair as [125]

$$r(\{\mathrm{y}(k)\}, \{\mathrm{x}(k)\}) \triangleq \lim_{k\to\infty} \sup \frac{1}{k} \mathrm{E}\left[L_k(\mathrm{x}_1^\infty)\right]. \tag{6.5}$$

In the sequel, we focus only on the MSE as the distortion metric and on Gaussian stationary process sources. Accordingly, we define the *average distortion* associated with an ED pair as:

$$d(\{\mathrm{x}(k)\},\{\mathrm{y}(k)\}) \triangleq \lim_{k\to\infty} \sup \frac{1}{k}\mathrm{E}\left[\|\mathrm{x}_1^k - \mathrm{y}_1^k\|^2\right]. \tag{6.6}$$

The notions of average operational rate and average distortion allow us to define the operational causal rate-distortion function as follows.

**Definition 6.2** (Operational Causal Rate-Distortion)**.** *The operational causal rate-distortion function for a source* $\{\mathrm{x}(k)\}$ *is defined as [125]:*

$$R_c^{op}(D) \triangleq \inf_{\substack{\{f_k\}\ \textit{causal},\\ d(\{\mathrm{x}(k)\},\{\mathrm{y}(k)\})\le D}} r(\{f_k\},\{\mathrm{x}(k)\}). \tag{6.7}$$

▲

We note that the operational causal rate distortion function defined above corresponds to the *optimal theoretically attainable performance* (OPTA) of any causal ED pair.

In order to define an information theoretical counterpart of $R_c^{op}(D)$, we notice from [63, Theorem 5.4.2] that

$$\frac{1}{k}\mathrm{E}\left[L_k(\mathrm{x}_1^\infty)\right] \ge \frac{1}{k}H(\mathbf{b}(k)), \quad \forall k \in \mathbb{Z}^+. \tag{6.8}$$

Also, from the Data Processing Inequality (see Fact 2.5 on page 40 of Chapter 2), we obtain

$$H(\mathbf{b}(k)) = I(\mathbf{b}(k);\mathbf{b}(k)) \ge I(\mathrm{x}_1^\infty;\mathrm{y}_1^k) = I(\mathrm{x}_1^k;\mathrm{y}_1^k), \tag{6.9}$$

where the last equality follows from the fact that, for a causal ED, (6.4) needs to hold. Thus, combining (6.5), (6.8) and (6.9),

$$r(\{\mathrm{y}(k)\},\{\mathrm{x}(k)\}) \ge \lim_{k\to\infty} \frac{1}{k} I(\mathrm{x}_1^k;\mathrm{y}_1^k) = \bar{I}(\{\mathrm{x}(k)\};\{\mathrm{y}(k)\}). \tag{6.10}$$

This lower bound motivates the introduction of an information-theoretic (as opposed to operational) causal rate distortion function, as defined below.

**Definition 6.3** (Information-Theoretic Causal Rate-Distortion Function)**.** *The information-theoretic causal rate-distortion function for a source* $\{\mathrm{x}(k)\}$*, with respect to the MSE distortion metric, is defined as*

$$R_c^{it}(D) \triangleq \inf \bar{I}(\{\mathrm{x}(k)\};\{\mathrm{y}(k)\}), \tag{6.11}$$

*where the infimum is over all processes* $\{\mathrm{y}(k)\}$ *such that* $d(\{\mathrm{x}(k)\},\{\mathrm{y}(k)\}) \le D$ *and such that* (6.4) *holds.* ▲

The above definition is a special case of the information-theoretic rate distortion function with delay introduced by Pinsker and Gorbunov in [169], which was then shown to converge to Shannon's RDF, for Gaussian stationary sources and in the limit as the rate goes to infinity [170] .

Since an ED pair matches Definition 6.1 if and only if its output $\{\mathrm{y}(k)\}$ satisfies (6.4) when the input is $\{\mathrm{x}(k)\}$, it follows from (6.7) and (6.10) that

$$R_c^{op}(D) \geq R_c^{it}(D). \tag{6.12}$$

It is known that the mutual information across an AWGN channel introducing noise with variance $D$, say $R_{AWGN}(D)$, exceeds Shannon's rate-distortion function $R(D)$ by at most $0.5$ bits/sample, see, e.g. [126]. Thus, we have:

$$R_c^{it}(D) \leq R_{AWGN}(D) \leq R(D) + 0.5. \tag{6.13}$$

This performance gap is consistent with the results reported in [171], where the gain produced by allowing for non-causal reconstruction in DPCM converters was found to be a MSE reduction of at most $3$ [dB].

In the sequel, we propose an iterative procedure to obtain an upper bound for $R_c^{it}(D)$ for Gaussian stationary process sources. This bound can be defined as follows:

**Definition 6.4** (Information-Theoretic Causal Stationary RDF)**.** *The Information-Theoretic Causal Stationary Rate-Distortion function $\overline{R_c^{it}}(D)$ is defined as*

$$\overline{R_c^{it}}(D) \triangleq \inf \bar{I}(\{\mathrm{x}(k)\}\,;\{\mathrm{y}(k)\}), \tag{6.14}$$

*where the infimum is over all processes $\{\mathrm{y}(k)\}$ such that:*

*i)* $d(\{\mathrm{x}(k)\}\,,\{\mathrm{y}(k)\}) \leq D$,

*ii) the reconstruction error $\{\mathrm{z}(k)\} \triangleq \{\mathrm{y}(k)\} - \{\mathrm{x}(k)\}$ is jointly stationary with the source, and*

*iii) Markov chain* (6.4) *holds.*

▲

We also find below that an upper bounding function for $R_c^{op}(D)$ can also be obtained from this iterative procedure, by showing that, for Gaussian stationary sources, it holds that

$$R_c^{op}(D) \leq \overline{R_c^{it}}(D) + 0.254, \quad \forall D > 0. \tag{6.15}$$

From Definition 6.4, it follows that $\overline{R_c^{it}}(D)$ is a tighter upper bound on $R_c^{it}$ than $R_{AWGN}(D)$, for all distortions. To see this, recall that

$$R_{AWGN}(D) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log_2\left(\frac{S_{\mathrm{x}}(\mathrm{e}^{j\omega})}{D}+1\right) d\omega, \quad 0 < D < \sigma_{\mathrm{x}}^2. \tag{6.16}$$

On the other hand, simply placing an optimal scalar gain, with the value $\sigma_{\mathrm{x}}^2/(\sigma_{\mathrm{x}}^2+d)$, after the output of the AWGN channel with variance $d$, the distortion is reduced to $d' \triangleq \sigma_{\mathrm{x}}^2 d/(\sigma_{\mathrm{x}}^2+d)$ or, equivalently,

$$d = \frac{\sigma_{\mathrm{x}}^2}{\sigma_{\mathrm{x}}^2 - d'} d'. \tag{6.17}$$

Notice that applying a non-zero scalar gain after the ED pair preserves the mutual information rate between source and reconstruction. Thus, the mutual information rate of an AWGN channel with the optimal scalar gain at the decoder end, as a function of the distortion, say $R'_{AWGN}(D)$, is given by $R'_{AWGN}(D) = R_{AWGN}(\frac{\sigma_{\mathrm{x}}^2}{\sigma_{\mathrm{x}}^2-D}D)$. Since $R_{AWGN}(D)$ is monotonically decreasing $\forall D \leq \sigma_{\mathrm{x}}^2$ (see (6.16)), it follows that $R'_{AWGN}(D) < R_{AWGN}(D)$. From this and (6.13), and noting that $\overline{R_c^{it}} \leq R'_{AWGN}(D)$, we conclude that

$$\overline{R_c^{it}}(D) < R(D) + 0.5 \text{ bits/sample}. \tag{6.18}$$

From Definition 6.4, it is also clear that, if there exists a realization of $R_c^{it}(D)$ in which the reconstruction error is jointly stationary with the source (which seems to be a reasonable conjecture), then $\overline{R_c^{it}}(D)$ actually coincides with $R_c^{it}(D)$.

## 6.2 Obtaining the Stationary Causal RDF

Here we show that, for Gaussian stationary sources, the stationary causal RDF $\overline{R_c^{it}}(D)$, introduced in Definition 6.4, can always be obtained by iteration. More specifically, we propose an iterative procedure which is guaranteed to converge to the causal stationary RDF. In addition, this procedure yields a characterization of the filters in a feedback quantizer that achieve an operational rate that equals the upper bound on the right hand side of (6.15).

To derive these results, we first consider a scheme consisting of an AWGN channel and a set of causal filters, as depicted in Fig. 6.2. In this scheme, the source $\{\mathrm{x}(k)\}$ is Gaussian and stationary, with PSD $S_{\mathrm{x}}(\mathrm{e}^{j\omega})$. From this, we define, as in Chapter 3,

$$\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega}) \triangleq \sqrt{S_{\mathrm{u}}(\mathrm{e}^{j\omega})}, \quad \forall \omega \in [-\pi, \pi]. \tag{6.19}$$

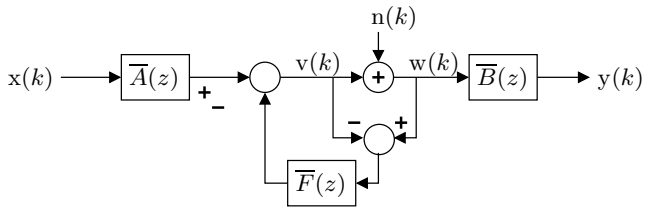



**Figure 6.2:** AWGN channel with causal filters.

In Fig. 6.3, $\{\mathrm{n}(k)\}$ is Gaussian noise with i.i.d. samples, independent of $\{\mathrm{x}(k)\}$. Thus, between $\mathrm{v}(k)$ and $\mathrm{w}(k)$ lies the AWGN channel $\mathrm{w}(k) = \mathrm{v}(k) + \mathrm{n}(k)$. The filters $\overline{A}(z)$ and $\overline{B}(z)$ are casual and stable. The filter $\overline{F}(z)$ is stable and strictly causal.

As in Chapter 3, we define

$$K \triangleq \frac{\sigma_{\mathrm{v}}^2}{\sigma_{\mathrm{n}}^2} + 1 = \frac{\sigma_{\mathrm{w}}^2}{\sigma_{\mathrm{n}}^2}, \tag{6.20}$$

$$f(\mathrm{e}^{j\omega}) \triangleq \left|1 - \overline{F}(\mathrm{e}^{j\omega})\right|, \quad \forall \omega \in [-\pi, \pi]. \tag{6.21}$$

The signal transfer function and the PSD of source uncorrelated distortion for the system in Fig. 6.2 are given respectively by

$$\widetilde{W}(z) \triangleq \overline{A}(z)\overline{B}(z), \tag{6.22a}$$

$$S_{\mathrm{u}}(\mathrm{e}^{j\omega}) \triangleq \left|\overline{B}(\mathrm{e}^{j\omega})\right|^2 f(\mathrm{e}^{j\omega})^2 \sigma_{\mathrm{n}}^2. \tag{6.22b}$$

In turn, the PSD of $\{\mathrm{w}(k)\}$ is given by

$$S_{\mathrm{w}}(\mathrm{e}^{j\omega}) = \Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})^2 \left|\overline{A}(\mathrm{e}^{j\omega})\right|^2 + \sigma_{\mathrm{n}}^2 f(\mathrm{e}^{j\omega})^2, \quad \forall \omega \in [-\pi, \pi]. \tag{6.22c}$$

From (3.23) (see page 53), we obtain that the MSE is

$$MSE = D_c \triangleq \frac{\|\Omega_{\mathrm{x}}\overline{A}\|^2 \|\overline{B}f\|^2}{K - \|f\|^2} + \|(\overline{A}\,\overline{B} - 1)\Omega_{\mathrm{x}}\|^2 \tag{6.23}$$

From this, we define the following

**Optimization Problem 6.1.** *For any given $\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})$, and for any given $K > 1$, find the causal filters $\overline{A}(z)$, $\overline{B}(z)$ and $\overline{F}(z)$ that minimize $D_c$.* ▲

Based upon the results obtained in Chapter 5, we next show that solving Optimization Problem 6.1 amounts to finding the stationary causal rate distortion function of Definition 6.4.

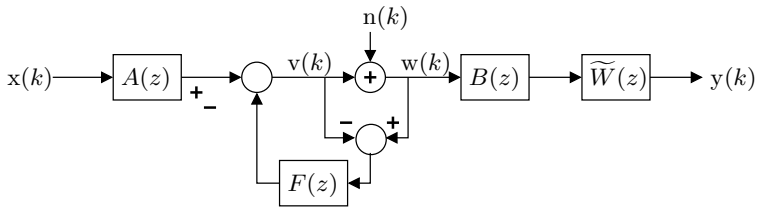



**Figure 6.3:** AWGN channel within a "perfect reconstruction" system and causal de-noising filter.

**Lemma 6.1.** *If the filters $\overline{A}^{\star}(z)$, $\overline{B}^{\star}(z)$, and $\overline{F}^{\star}(z)$ solve Optimization Problem 6.1 and yield distortion $D_c^{\star}$, then*

$$\frac{1}{2}\ln(K) = \overline{R_c^{it}}(D_c^{\star}). \tag{6.24}$$

▲

*Proof.* Let $S'_u$ and $W'$ be, respectively, the PSD of the source uncorrelated noise, and the signal transfer function of a realization of $\overline{R_c^{it}}(D)$ (see (6.22)). Then

$$\frac{1}{2}\ln(K) = I(\mathrm{v}(k);\mathrm{w}(k)) \overset{(a)}{\geq} \bar{I}(\{\mathrm{v}(k)\}\to\{\mathrm{w}(k)\}) \overset{(b)}{\geq} \bar{I}(\{\mathrm{x}(k)\};\{\mathrm{y}(k)\}) \overset{(c)}{\geq} \overline{R_c^{it}}(D_c). \tag{6.25}$$

In the above, the first equality follows from the fact that $\{\mathrm{n}(k)\}$ is Gaussian. Inequalities $(a)$ and $(b)$ follow directly from Lemma 5.1. Lemma 5.1 also shows that equality is achieved in $(a)$ if and only if $\{\mathrm{w}(k)\}$ is white, and in $(b)$ if and only if $\mathcal{N}_B \subseteq \mathcal{N}_A$, i.e., if and only if $B(z)$ is invertible for all frequencies $\omega$ for which $\left|A(\mathrm{e}^{j\omega})\right| > 0$. Inequality $(c)$ follows from the definition of $\overline{R_c^{it}}(D)$. From Lemma 4.1, the reconstruction error that realizes $\overline{R_c^{it}}(D)$ needs to be Gaussian. Since, in the system of Fig. 6.2, the distortion is Gaussian, equality is achieved in $(c)$ iff

$$S_{\mathrm{u}}(\mathrm{e}^{j\omega}) = S'_{\mathrm{u}}(\mathrm{e}^{j\omega}), \quad \forall\omega\in[-\pi,\pi], \text{ and} \tag{6.26a}$$

$$\widetilde{W}(\mathrm{e}^{j\omega}) = \widetilde{W}'(\mathrm{e}^{j\omega}), \quad \forall\omega\in[-\pi,\pi]. \tag{6.26b}$$

We note that, despite the causality constraints on the filters, (6.26) can be met while yielding $S_{\mathrm{w}}(\mathrm{e}^{j\omega})$ constant, for any $\Omega_{\mathrm{x}}$, see (6.22). Thus, $\overline{A}^{\star}(z)$, $\overline{B}^{\star}(z)$, and $\overline{F}^{\star}(z)$ solve Optimization Problem 6.1 if and only if they yield $\frac{1}{2}\ln(K) = \overline{R_c^{it}}(D_c)$. This completes the proof. □

For any choice of filters $\overline{A}(z)$, $\overline{B}(z)$, and $\overline{F}(z)$, the system in Fig. 6.2 is equivalent to the one depicted in Fig. 6.3. In Fig. 6.3,

$$A(z) = \overline{A}(z), \tag{6.27a}$$

$$F(z) = \overline{F}(z), \tag{6.27b}$$

$$B(z) = \overline{A}(z)^{-1}, \text{ and} \tag{6.27c}$$

$$\widetilde{W}(z) = \overline{B}(z)B(z)^{-1} = \overline{B}(z)A(z). \tag{6.27d}$$

Thus, $A(z)$ and $B(z)$ satisfy the perfect reconstruction condition

$$A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega}) \equiv 1, \tag{6.28}$$

and $\widetilde{W}(z)$ is the signal transfer function of the system, as before. On the other hand, the net effect of the AWGN channel and the filters $A(z)$, $B(z)$ and $F(z)$ is to introduce Gaussian stationary additive noise, independent of the source. We also have that

$$\mathrm{y}(k) = \widetilde{W}(z)\,\mathrm{x}(k) + \widetilde{W}B(z)(z)[1 - F(z)]\,\mathrm{n}(k). \tag{6.29}$$

From this, the PSD of source uncorrelated noise, $S_{\mathrm{u}}(\mathrm{e}^{j\omega})$, is given by

$$S_{\mathrm{u}}(\mathrm{e}^{j\omega}) = \left|\widetilde{W}(\mathrm{e}^{j\omega})\right|^2 \left|B(\mathrm{e}^{j\omega})\right|^2 \left|1 - F(\mathrm{e}^{j\omega})\right|^2 \sigma_{\mathrm{n}}^2. \tag{6.30}$$

Thus, $\widetilde{W}(z)$ can be seen as a de-noising filter utilized to reduce the MSE of a Gaussian stationary source $\{\mathrm{x}(k)\}$ corrupted by additive Gaussian stationary noise with PSD $S_{\mathrm{u}}(\mathrm{e}^{j\omega})$. Substituting (6.27) into (6.23), the MSE can be expressed as

$$D_c = \sigma_{\mathrm{u}}^2 + \|(\widetilde{W} - 1)\Omega_{\mathrm{x}}\|^2 = \frac{\|\Omega_{\mathrm{x}}A\|^2\|\widetilde{W}Bf\|^2}{K - \|f\|^2} + \|(\widetilde{W} - 1)\Omega_{\mathrm{x}}\|^2, \tag{6.31}$$

where $\sigma_{\mathrm{u}}^2$ is the variance of the source uncorrelated reconstruction error.

In addition to (6.28), for any given $F(z)$ and $\widetilde{W}(z)$, the filters $A(z)$ and $B(z)$ in Fig. 6.3 are chosen so as to minimize the variance of source uncorrelated noise. For this purpose, from the viewpoint of the subsystem comprised of the AWGN channel and the filters $A(z)$, $B(z)$ and $F(z)$, the filter $\widetilde{W}(z)$ acts as a frequency weighting filter. Thus, for any $F(z)$ and $\widetilde{W}(z)$, the filters $A(z)$ and $B(z)$ that minimize $\sigma_{\mathrm{u}}^2$ can be found from Theorem 3.3 (see page 56), by setting $a = b = 1$, $W(\mathrm{e}^{j\omega}) \triangleq 1$, and $P(z) = \widetilde{W}(z)$. This yields that $A(z)$ and $B(z)$ satisfy

$$\left|A(\mathrm{e}^{j\omega})\right| = \kappa\sqrt{|P(\mathrm{e}^{j\omega})|\,|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})|^{\sim 1}\,f(\mathrm{e}^{j\omega})\,|W(\mathrm{e}^{j\omega})|}, \quad \text{a.e. on } [-\pi, \pi], \tag{6.32a}$$

$$\left|B(\mathrm{e}^{j\omega})\right| = \frac{1}{\kappa}\sqrt{|P(\mathrm{e}^{j\omega})|^{\sim 1}\,|\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})|\,f(\mathrm{e}^{j\omega})^{\sim 1}\,|W(\mathrm{e}^{j\omega})|}, \quad \text{a.e. on } [-\pi, \pi], \tag{6.32b}$$

where $\kappa > 0$ is an arbitrary real constant. Also, from (3.42) (page 56), the variance of source uncorrelated error when (6.32) holds is given by

$$\sigma_{\mathrm{u}}^2 = \frac{\langle\Omega_{\mathrm{x}}|\widetilde{W}|, f\rangle^2}{K - \|f\|^2} \tag{6.33}$$

On the other hand, the filter $F(z)$ needs to be such that

$$\int_{-\pi}^{\pi} \log f(\mathrm{e}^{j\omega}) d\omega \geq 0, \tag{6.34}$$

see (3.76) on page 64.

Thus, if one wishes to minimize the reconstruction MSE by choosing appropriate *causal* filters in the system in Fig. 6.3 for a given value of $K$, one needs to solve the following optimization problem.

**Optimization Problem 6.2.** *For any given $\Omega_{\mathrm{x}}(\mathrm{e}^{j\omega})$, and for any given $K > 1$, find the frequency response $\widetilde{W}(\mathrm{e}^{j\omega})$ and the frequency response magnitude $f(\mathrm{e}^{j\omega})$ that*

$$\textit{Minimize:}\quad D_c \triangleq \frac{\langle \Omega_{\mathrm{x}} |\widetilde{W}|, f\rangle^2}{K - \|f\|^2} + \|(\widetilde{W} - 1)\Omega_{\mathrm{x}}\|^2 \tag{6.35a}$$

$$\textit{Subject to:}\quad \widetilde{W} \in \mathbb{H}, \tag{6.35b}$$

$$\int_{-\pi}^{\pi} \ln f(\mathrm{e}^{j\omega}) d\omega \geq 0, \tag{6.35c}$$

*where $\mathbb{H} \subset \mathbb{G}$ denotes the space of all frequency responses that can be obtained with causal filters.* ▲

Recalling that the system in Fig. 6.2 can always be re-arranged in the form of the system in Fig. 6.3 with filters satisfying (6.27), it becomes clear that Optimization Problem 6.2 is equivalent to Optimization Problem 6.1. We put this fact in the form of a lemma for future reference.

**Lemma 6.2.** *For any given $\Omega_{\mathrm{x}}$ and $K > 1$, Optimization Problem 6.2 is equivalent to Optimization Problem 6.1.*

The advantages for the analysis that the system of Fig. 6.3 has over the system of Fig. 6.2 will become evident after we state the following lemma, which is key for subsequent results in this chapter. It will play a central role in demonstrating the convergence properties of the iterative procedure that yields $\overline{R_c^{it}}(D)$, to be proposed later.

**Lemma 6.3.** *Define the sets of functions*

$$\mathbb{F}_K \triangleq \left\{ f : [-\pi, \pi] \to \mathbb{R}_0^+, \|f\|^2 < K \right\}, \tag{6.36}$$

$$\mathbb{G} \triangleq \left\{ G : [-\pi, \pi] \to \mathbb{C} \right\}, \tag{6.37}$$

*where $K$ is some positive constant. Then, for any $G \in \mathbb{W}$ and $K > 1$, the cost functional $\mathscr{J} : \mathbb{F}_K \times \mathbb{G} \to \mathbb{R}_0^+$, defined as*

$$\mathscr{J}(f, g) \triangleq \frac{\langle f, |g|\rangle^2}{K - \|f\|^2} + \|g - G\|^2, \tag{6.38}$$

*is convex.* ▲

*Proof.* Choose any two arbitrary pairs $(f_1, g_1)$ and $(f_2, g_2)$, and a third arbitrary pair $(f_0, g_0)$ satisfying

$$(f_0, g_0) \triangleq \lambda(f_1, g_1) + [1-\lambda](f_2, g_2)$$

for some $\lambda \in [0, 1]$. Defining

$$\eta \triangleq f_2 - f_1; \quad \theta \triangleq g_2 - g_1,$$

any duplet along the "line" between $(f_1, g_1)$ and $(f_2, g_2)$ can be written in terms of a single parameter $s$ via

$$(f, g) = (f_0 + \eta s\,,\, g_0 + \theta s),$$

where $s \in [\lambda - 1, \lambda]$. Substitution into (6.38) yields

$$\mathscr{J}(f, g) = J(s) \triangleq \frac{\left(a + bs + ds^2\right)^2}{D} + R + es + \|\theta\|^2 s^2 \tag{6.39}$$

where

$$\begin{aligned}
a &\triangleq \langle f_0, |g_0| \rangle && (6.40)\\
b &\triangleq \langle f_0, |\theta| \rangle + \langle |g_0|\,, \eta \rangle && (6.41)\\
d &\triangleq \langle \eta, |\theta| \rangle && (6.42)\\
D &\triangleq K - \|f_0\|^2 - 2\langle f_0, \eta \rangle s - \|\eta\|^2 s^2 && (6.43)\\
e &\triangleq 2\,\mathcal{R}\left\{\langle g_0 - G,\, \theta \rangle\right\} && (6.44)\\
R &\triangleq \|g_0\|^2 + \|G\|^2 - 2\mathcal{R}\left\{\langle g_0, G \rangle\right\}, && (6.45)
\end{aligned}$$

where $\mathcal{R}\{x\}$ denotes the real part of $x$. We next show that $\mathscr{J}(\cdot, \cdot)$ is convex along the line between $(f_1, g_1)$ and $(f_2, g_2)$. For this purpose, we first take the derivative of $J(s)$ with respect to $s$, yielding:

$$J'(s) = \frac{2\left(a + bs + ds^2\right)(b + 2ds)\, D - \left(a + bs + ds^2\right)^2 D'}{D^2} + e + 2\|\theta\|^2 s.$$

Differentiating again and evaluating all terms at $s = 0$, we obtain

$$\begin{aligned}
J''(0) &= \frac{\left\{2(b^2 + 2ad)D_0 + 2abD_0' - 2abD_0' - a^2 D_0''\right\} D_0^2 - 2(2abD_0 - a^2 D_0')D_0 D_0'}{D_0^4} + 2\|\theta\|^2 \\
&= \frac{2b^2 D_0^2 + 4adD_0^2 - a^2 D_0'' D_0 - 4abD_0 D_0' + 2a^2 (D_0')^2}{D_0^3} + 2\|\theta\|^2 \\
&= \frac{\frac{2}{D_0}\left(bD_0 - aD_0'\right)^2 + 4adD_0 - a^2 D_0'' + 2\|\theta\|^2 D_0^2}{D_0^2}, && (6.46)
\end{aligned}$$

where

$$D_0 \triangleq D\big|_{s=0} = K - \|f_0\|^2 \tag{6.47}$$

$$D_0' \triangleq \frac{\partial D}{\partial s}\Big|_{s=0} = -2\langle f_0, \eta\rangle \tag{6.48}$$

$$D_0'' \triangleq \frac{\partial^2 D}{\partial s^2}\Big|_{s=0} = -2\|\eta\|^2. \tag{6.49}$$

Substituting (6.49) and (6.42) into (6.46),

$$\begin{aligned} J''(0) &= \frac{\frac{1}{D_0}\left(bD_0 - aD_0'\right)^2 + 2a\langle \eta, |\theta|\rangle D_0 + a^2\|\eta\|^2 + \|\theta\|^2 D_0^2}{D_0^2/2} \\ &= \frac{\frac{1}{D_0}\left(bD_0 - aD_0'\right)^2 + \|\eta a + |\theta|\, D_0\|^2}{D_0^2/2} \geq 0. \end{aligned}$$

Thus, the cost $\mathscr{J}(\cdot,\cdot)$ along the "line" between $(f_1, g_1)$ and $(f_2, g_2)$ is convex. Since the latter holds for any arbitrary pair of pairs, if follows that $\mathscr{J}(f, g)$ is convex. This completes the proof. □

**Lemma 6.4.** *For all $\Omega_{\mathrm{x}}$ and for all $K > 1$, Optimization Problem 6.2 is convex .* ▲

*Proof.* With the change of variables $G \triangleq \Omega_{\mathrm{x}}$ and $g \triangleq \Omega_{\mathrm{x}}\widetilde{W}$, we obtain $D_c = \mathscr{J}(f, g)$. With this, Optimization Problem 6.2 amounts to finding the functions $f$ and $g$ that

$$\text{Minimize: } \mathscr{J}(f, g) \tag{6.50a}$$

$$\text{Subject to: } g \in \mathbb{W}, f \in \mathbb{B}. \tag{6.50b}$$

where

$$\mathbb{W} \triangleq \{g = \Omega_{\mathrm{x}} W : W \in \mathbb{H}\} \tag{6.51}$$

$$\mathbb{B} \triangleq \left\{ f \in \mathbb{F}_K : \int_{-\pi}^{\pi} \ln f(\mathrm{e}^{j\omega}) d\omega = 0 \right\}. \tag{6.52}$$

Clearly, $\mathbb{H}$ is a convex set. This implies that $\mathbb{W}$ is a convex set. In addition, $\mathbb{B}$ is also a convex set, and from Lemma 6.3, $\mathscr{J}(f, g)$ is a convex functional. Therefore, the optimization problem stated in (6.50) is convex. This implies that Optimization Problem 6.2 is convex, thus completing the proof. □

We can now define an iterative procedure that, as will be shown later, yields the information theoretic causal rate distortion function:

***Iterative Procedure 1***

*For any target information theoretical rate $R$,*

*Step 1: Set $K = 2^{2R}$.*

*Step 2: Set $\widetilde{W}(\mathrm{e}^{j\omega}) \equiv 1$.*

*Step 3: Find the frequency response magnitude $f \in \mathbb{B}$ that minimizes $D_c$ for given $\widetilde{W}$.*

*Step 4: Find the causal frequency response $\widetilde{W} \in \mathbb{H}$ that minimizes $D_c$ for given $f$.*

*Step 5: Return to step 3.*

Notice that after solving Step 3 in the first iteration of Iterative Procedure 1, the result is $R^{\perp}(D)$, i.e., the MSE is comprised of only source-uncorrelated distortion. Step 4 then reduces the MSE by attenuating source-uncorrelated noise at the expense of introducing linear distortion. Each step reduces the MSE until a local (or global) minimum of the MSE is obtained. Based upon the convexity of Optimization Problem6.2, the following theorem, which is the main technical result in this chapter, guarantees convergence to the global minimum of the MSE for a given end-to-end mutual information. Since all the filters are causal, this global minimum actually corresponds to a point on the $\overline{R_c^{it}}(D)$ plot.

**Theorem 6.5** (Convergence of Iterative Procedure 1)**.** *Iterative Procedure 1 converges to the unique $f$ and $\widetilde{W}$ that realize $R_c^{it}(D)$. More precisely, if the MSE obtained by the procedure for a rate $R$, in the limit as the number of iterations tends to infinity, is $D'$, then we have $R = \overline{R_c^{it}}(D')$.* ▲

*Proof.* The result follows directly from the fact that Optimization Problem 6.2 is convex in $f$ and $\widetilde{W}$, which was shown in Lemma 6.3, and from Lemmas 6.2 and 6.1. □

The above theorem states that the stationary information theoretic causal rate-distortion function can be obtained by using Iterative Procedure 1. In practice, this means that an approximation arbitrarily close to $R_c(D)$ for a given $D$ can be obtained if sufficient iterations of the procedure are carried out.

The feasibility of running Iterative Procedure 1 depends on the feasibility of solving each of the minimization sub-problems involved in steps 3 and 4. We next show how these sub-problems can be solved.

### Solving Step 3

If $\widetilde{W}(\mathrm{e}^{j\omega})$ is given, the minimization problem in Step 3 of Iterative Procedure 1 is equivalent to solving a feedback quantizer design problem with the constraint $A(z)B(z) \equiv 1$ and with error weighting filter $P(\mathrm{e}^{j\omega}) = \widetilde{W}(\mathrm{e}^{j\omega})$. Therefore, the solution is given by Theorem 3.10, with the choice $a = 1$, $b = \infty$, and setting $P(z) = \widetilde{W}(z)$.

### Solving Step 4

Finding the causal frequency response $\widetilde{W}(\mathrm{e}^{j\omega}) \in \mathbb{H}$ that minimizes $D_c$ for a given $f$ is equivalent to solving

$$\min_{g:g\in\mathbb{W}} \mathscr{J}(f,g) \tag{6.53}$$

for a given $f$, where $\mathbb{W}$ is as defined in (6.51). Since $\mathbb{W}$ and $\mathscr{J}(\cdot,\cdot)$ are convex, (6.53) is a convex optimization problem. As such, its global solution can always be found iteratively. In particular, if $\widetilde{W}(z)$ is constrained to be an $M$-th order FIR filter with impulse response $\boldsymbol{c} \in \mathbb{R}^{M+1}$, such that $\widetilde{W}(\mathrm{e}^{j\omega}) = \mathcal{F}\{\boldsymbol{c}\}$, where $\mathcal{F}\{\cdot\}$ denotes the discrete-time Fourier transform, then

$$\mathscr{G}(\boldsymbol{c}) \triangleq \mathscr{J}(f, \mathcal{F}\{\boldsymbol{c}\}) \tag{6.54}$$

is a convex functional. The latter follows directly from the convexity of $\mathscr{J}(\cdot,\cdot)$ and from Lemma 6.6 (see page 206). As a consequence, one can solve the minimization problem in Step 4, to any degree of accuracy, by minimizing $\mathscr{G}(\boldsymbol{c})$ over the values of the impulse response of $\widetilde{W}(\mathrm{e}^{j\omega})$, using standard convex optimization methods (see, e.g, [147]). This approach also has the benefit of being amenable to numerical computation.

## 6.3 Upper Bound on the Operational Causal RDF

By using entropy coded scalar quantization with dither, the operational rate of an FQ with the filters obtained via Iterative Procedure 1 is guaranteed to exceed $\overline{R_c^{it}}(D)$ by less than $0.254$ bits/sample, see Remark 4.5 after Lemma 4.10 on page 148. Thus, we have the bound

$$R_c^{op}(D) \leq \overline{R_c^{it}}(D) + 0.254 \quad \text{bits/sample.} \tag{6.55}$$

We note that the feedback quantizer thus obtained corresponds to the ED pair yielding the best operational rate-distortion performance achievable by any ED pair that uses only LTI filters and subtractively dithered scalar quantization.

If the requirement of zero-delay, which is stronger than that of causality, was to be satisfied, then it would not be possible to apply entropy coding to long sequences of quantized samples. This would entail an excess bit-rate not greater than 1 bit per sample, see, e.g., [63, Section 5.4]. As a consequence of this, the upper bound on the operational bit-rate with zero-delay, say $R_{ZD}^{op}(D)$, would be

$$R_{ZD}^{op}(D) \leq \overline{R_c^{it}}(D) + 0.254 + 1 \quad \text{bits/sample.} \tag{6.56}$$

The $0.254$ bits at the end of (6.55), commonly referred to as the "space-filling loss" of scalar quantization, can be reduced by using vector quantization [65, 126]. Vector quantization could be applied while preserving causality (and without introducing delay) if the samples of the source were $N$-dimensional vectors. This would also allow for the use of entropy coding over $N$-dimensional vectors of quantized samples, which reduces the extra 1 bit/sample at the end of (6.56) to $1/N$ bit/sample, see [63, Theorem 5.4.2].

## 6.4 Summary

In this chapter we have shown that an upper bound on the information-theoretic causal rate distortion function for Gaussian stationary sources and MSE distortion criterion, denoted by $R_c^{it}$, can always be found iteratively. For that purpose, we have introduced an iterative algorithm which converges to the minimum mutual information rate between source and reconstruction achievable by any stationary error process having a given variance $D$. We have named the associated minimum as the stationary causal rate distortion function, denoted by $\overline{R_c^{it}}(D)$. If there exists a realization of the causal RDF for Gaussian stationary sources and MSE distortion metric in which the reconstruction error is jointly stationary with the source, then $\overline{R_c^{it}}(D) = R_c^{it}(D)$. The proposed method also yields the frequency response of the filters in a feedback quantizer, using entropy coded scalar quantization with subtractive dither, with which the operational rate exceeds $\overline{R_c^{it}}(D)$ by at most $0.254$ bits/sample. This constitutes an upper bound to the operational rate of any causal encoder-decoder pair.

## 6.5 Appendix

**Lemma 6.6.** *Let $\mathscr{C} : \mathbb{H} \to \mathbb{R}$ be a convex cost functional. Let $\mathcal{T} : \mathbb{X} \to \mathbb{H}$ be a linear mapping, where $\mathbb{X}$ is some given vector space. Then, the functional*

$$\mathscr{G}(x) \triangleq \mathscr{C}(\mathcal{T}x + b) \tag{6.57}$$

*is convex.*

*Proof.* Let $x_1$, $x_2$ be any two vectors in $\mathbb{X}$. For any scalar parameter $\lambda \in [0, 1]$,

$$\begin{aligned}
\mathscr{G}(\lambda x_1 + [1-\lambda]x_2) &= \mathscr{C}(\mathcal{T}(\lambda x_1 + [1-\lambda]x_2) + b) && (6.58)\\
&\overset{(a)}{=} \mathscr{C}(\lambda(\mathcal{T}x_1 + b) + [1-\lambda](\mathcal{T}x_2 + b)) && (6.59)\\
&\overset{(b)}{\leq} \lambda\mathscr{C}(\mathcal{T}x_1 + b) + [1-\lambda]\mathscr{C}(\mathcal{T}x_2 + b) && (6.60)\\
&= \lambda\mathscr{G}(x_1) + [1-\lambda]\mathscr{G}(x_2), && (6.61)
\end{aligned}$$

where $(a)$ stems from the linearity of $\mathcal{T}$ and $(b)$ follows from the fact that $\mathscr{C}(\cdot)$ is convex. This completes the proof. □

# Chapter 7

# Conclusions

*Previously I did not understand why I got no answer to my question;*
*today I do not understand how I could believe I was capable of asking.*
*But I didn't really believe, I only asked.*
*Franz Kafka, Bohemian novelist.*

*Each problem that I solved became a rule,*
*which served afterwards to solve other problems.*
*René Descartes, French Philosopher, in "Le Discours de la Méthode".*

## 7.1 Overview

This thesis has presented several novel results on the performance and design of both entropy and resolution constrained coders and decoders for stochastic sources. We next give a summary of the main contributions and point at directions of future research.

## 7.2 Main Contributions

We have introduced in Section 1.3.1 a new distortion metric, which extends the standard mean squared error (MSE). This extension has been given here the name *weighted correlation mean squared error* (WCMSE). This is the distortion metric considered throughout most of the thesis. The WCMSE is a weighted sum of two terms:

1. The first term is the component of the MSE which is uncorrelated to the source.

2. The second term is the remainder of the MSE.

By giving different weights, $a$ and $b$, to each of these terms, the WCMSE can take account of, for example, the different impact that each component of the MSE may have in some applications, such as image processing, parallel quantization schemes, and networked control systems. This is the first, and preliminary, contribution of this thesis.

The second contribution is the characterization of the filters around a scalar quantizer with given signal-to-noise ratio (SNR) that minimize the WCMSE. This was the subject of Chapter 3. These results were obtained by modelling quantization errors as white and uncorrelated with the source. This assumption is referred to as the *Linear Model*. The associated optimal performance (SNR-distortion) trade-off for this class of encoder-decoder pairs has also been established, and is summarized in Table 3.3 on page 87, for several architectural constraints. The rate-distortion performance of oversampled feedback quantization has also been analyzed in Section 3.12. In particular, we have shown that, for a fixed quantizer SNR, and when quantizer overload errors are negligible, the frequency weighted MSE of optimal perfect reconstruction feedback quantizers decreases exponentially with the oversampling ratio, $\lambda$. This result implies that, when entropy coded scalar quantization with subtractive dither and sufficient quantization levels to avoid overload is employed, the MSE can be made to decay with $\lambda$ as $\mathcal{O}(2^{-1.746\lambda})$, when $\lambda \to \infty$. We also obtained an extension of this result for the case of subtractively dithered scalar quantization with a (finite) number of quantization levels that is insufficient to avoid quantizer overload. For this case, it was shown that for a subtractively dithered scalar quantizer with $N$ levels, the MSE can be made to decay with $\lambda$ as $\mathcal{O}(\mathrm{e}^{-c_0\lambda^{1/3}})$, when $\lambda \to \infty$, where $c_0 \triangleq [0.5(N-1)]^{2/3}$. In order to achieve this asymptotic decay rate, it is necessary to balance the variance of clipping and granular errors in the quantizer, for each oversampling ratio, by adjusting the loading factor $\rho$ as $\rho = 4^{-1/3}\sqrt{3}(N-1)^{2/3}\lambda^{1/3}$. To the best of the author's knowledge, this is the only result available in the literature combining quantization with overload and oversampling. It also seems to be the first decay rate bound for the MSE of oversampled quantization that holds for sources with infinite support.

The third main contribution of this thesis was the characterization, in Chapter 4, of the rate-distortion function (RDF) for Gaussian sources wherein WCMSE is used as the distortion metric. We denoted this RDF by $R_{a,b}(D)$. First we showed that the WCMSE cannot be expressed as the expectation of a distortion measure in the usual sense, see Section 4.2.1. The case of scalar Gaussian sources was studied in Section 4.3. It was shown that, for scalar Gaussian sources, $R_{a,b}(D)$ is convex if and only if $a \leq 2b$. In Section 4.4 we characterized $R_{a,b}(D)$ for the case of vector Gaussian sources. This result allowed us to find $R_{a,b}(D)$ for stationary Gaussian sources in Section 4.5, and for Gaussian stationary vector process sources in Section 4.6. We studied special cases of $R_{a,b}(D)$ in Section 4.5.2. More specifically, it was verified that, as expected, $R_{1,1}(D) \equiv R(D)$, where $R(D)$ denotes Shannon's quadratic Gaussian

rate distortion function. Similarly, it was also verified that $R_{1,\infty}(D) = R^{\perp}(D)$, where the latter denotes the quadratic-Gaussian rate distortion function for source un-correlated distortions. In Section 4.9, we extended the characterization of $R^{\perp}(D)$ to cases in which there exists linear time invariant feedback from reconstruction to source. Part of the difficulty, and importance, associated with this result stems from the fact that we include the possibility of having unstable LTI filters in the feedback loop.

The fourth main contribution of this thesis, contained in Chapter 5, was the derivation of necessary and sufficient conditions under which the optimal filters for SNR constrained ED pairs, using scalar quantizers and the Linear Model, are such that they yield a realization of $R_{a,b}(D)$ when the scalar quantizer is replaced by an AWGN channel. The case of stationary processes was solved in Section 5.2. It was shown that, in this case, the sufficient conditions developed amount to having sufficient degrees of freedom to: i) whiten the output of the scalar quantizer; ii) yield the unique signal transfer function of a realization of $R_{a,b}(D)$, and iii) generate source-uncorrelated noise with the unique power spectral density required to realize $R_{a,b}(D)$. Sufficient conditions for vector sources were established in Section 5.3. In this case, the SNR constrained scheme corresponds to a transform coder with feedback and individual scalar quantizers in each subband. The SNR constraint can be any monotonically increasing function of $\frac{1}{2}\sum_k \log_2(\gamma(k)+1)$, where $\gamma(k)$ is the SNR associated with the quantizer in the $k$-th subband. For this case, the conditions can be summarized as having enough design freedom so that: i) the output signals of the scalar quantizers are uncorrelated with the outputs of the other scalar quantizers, ii) the signal transfer matrix must equal the unique matrix that realizes $R_{a,b}(D)$, and iii) the covariance matrix of the source uncorrelated reconstruction equals the unique additive-noise covariance matrix required to realize $R_{a,b}(D)$. Interestingly, these conditions imply that, under the Linear Model and when feedback is used, the signals that enter the scalar quantizers in an optimal causal transform coder must be correlated, at all rates. This conclusion departs from the situation‘ with non-causal transform coders, in which, under the Linear Model, analysis matrices that achieve total un-correlation of transform coefficients can be optimal, see, e.g., [58] and the references therein. Sufficient conditions for vector processes were derived in Section 5.4. The SNR constrained setting in this case corresponds to having a set of parallel scalar quantizers combined with a pre-filter matrix, an error feedback filter matrix, and a post-filter feedback matrix. The SNR constraint may take the form of any monotonically increasing function of $\frac{1}{2}\sum_k \log_2(\gamma_k+1)$, where $\gamma_k$ is the SNR associated with the quantizer in the $k$-th subband. It was shown there that these conditions amount to being able to: i) make the output of the scalar quantizers to have a diagonal and constant covariance matrix; ii) achieve an end-to-end signal transfer matrix that equals the unique signal transform matrix characterizing $R_{a,b}(D)$, and iii) yield a source-uncorrelated reconstruction error with a covariance matrix that equals the unique covariance matrix of source uncorrelated distortion required

to realize $R_{a,b}(D)$. When this scheme is associated with a filter bank, this result implies that, when feedback is available, an optimal filter bank does not need to satisfy either the un-correlation condition or the majorization condition. In all cases, part of the applicability of this result stems from the fact that it allows one to find the optimal filters, matrices, or filter matrices, for ED pairs that minimize distortion for a given operational bit-rate.

The last main contribution of this thesis, developed in Chapter 6, is the introduction of an iterative procedure which allows one to obtain upper bounds on the causal rate-distortion function for Gaussian stationary sources under the MSE distortion criterion. The bound obtained with this procedure, denoted by the function $\overline{R_c^{it}}(D)$, is tighter than $0.5$ bits per sample, at all rates. To the best of the author's knowledge, this is the tightest general bound for Gaussian stationary sources with memory available in the literature. Moreover, it was shown that, if there exists a realization of the causal rate distortion function, denoted by $R_c^{it}(D)$, in which the reconstruction error is jointly stationary with the source, then $\overline{R_c^{it}}(D) = R_c^{it}(D)$. The iterative procedure proposed here also yields a characterization of filters which, when employed in a feedback quantizer using entropy coded scalar quantization and subtractive dither, achieve an operational rate that exceeds $\overline{R_c^{it}}(D)$ by not more than $0.254$ bits/sample. This operational rate constitutes an upper bound on the minimum operational rate achievable by any causal source coder for Gaussian stationary sources and MSE distortion criterion.

## 7.3 Directions for Future Research

The results presented in this thesis are related to a number of related unsolved problems, opening the door to possible solutions. The following is a list of a few of these problems, some of which are already being considered by the author.

1. There are several optimal filter design problems under architectural constraints that haven't been treated in this thesis. A particularly challenging case is the one in which one can only measure the output of the scalar quantizer, but not inject signals after it, and one can only inject signals before the quantizer, but not measure signals before it. This situation is more restrictive than the design optimization problem solved in Section 3.6, where it was possible to inject and measure signals before quantization. The former problem is of practical importance, for example, when only one sensor is available (the one in the encoder), and where the transfer function from the quantized signal to the reconstructed signal has been fixed and cannot be altered.

2. The optimal filters characterized in Chapter 3 were not subject to complexity constraints such as filter order. In some cases, as it happens when optimizing the three filters, the expressions

obtained correspond to non-rational filters, which can be approximated arbitrarily well by using rational filters of sufficiently large order. It would be useful to obtain bounds on the performance degradation that would arise from imposing constraints on the order of the filters. Characterizing the optimal, un-restricted order filters first, and then approximating their characteristics with finite-order filters would provide a (top-down) design method for designing $\Sigma\Delta$ modulators, alternative to the (bottom-up) design methodologies usually described in the literature, see, e.g., [43, 45, 78, 172]. The merits of such method are yet to be determined.

3. The asymptotic decay rate of the reconstruction MSE with the oversampling ratio for dithered quantization with clipping obtained in Section 3.12.3 involved the use of several loose inequalities. This suggests that faster decay rates could be obtained.

4. A challenging and important direction of future research is the characterization of the WCMSE-RDF, with arbitrary weights, for situations in which there is LTI feedback between reconstruction and source, and where one or more transfer functions in the loop is unstable. Solving this problem would be an important step toward finding a solution to the open problem of optimal design of networked control loop systems under data-rate constraints.

5. It would be of practical interest to find the weights of the WCMSE that better represent perceived distortion in image processing applications. Once such weights are determined, it would be possible to design WCMSE-optimal image coders based on the results presented in Chapters 4 and 5. The perceived distortion-rate performance of such image coders could then be assessed by subjective or objective tests and compared to "state of the art" image compression methods.

6. As discussed in Chapter 6, the stationary causal RDF introduced in Definition 6.4 would correspond to the information-theoretic RDF if there exists a realization of the latter in which reconstruction error is jointly stationary with the source. The existence of such a realization seems a reasonable conjecture, which, to the best of the author's knowledge, has not been proven.

7. A refinement of the iterative procedure introduced in Chapter 6 (page 204) could be obtained if the convexity of the following optimization problem could be demonstrated: In relation to the scheme shown in Fig. 6.3, for a given SNR $\gamma = \sigma_{\mathrm{v}}^2/\sigma_{\mathrm{n}}^2$ and a given feedback filter $F(z)$, find the optimal *causal* filters $A(z)$, $B(z)$ and $\widetilde{W}(z)$. If the latter optimization problem is convex, then Step 4 in Iterative Procedure 1 could be carried out by repeating iteratively the following steps: a) first make $\widetilde{W}(z)$ be the causal Wiener filter for the source $\{\mathrm{x}(k)\}$ correlated by additive noise with PSD $\sigma_{\mathrm{n}}^2 \left|B(\mathrm{e}^{j\omega})\right|^2 \left|1 - F(\mathrm{e}^{j\omega})\right|^2$, then, b) use Theorem 3.5 to find the optimal $A(z)$ and $B(z)$. In

comparison with the method for solving Step 4 described on page 205, the former procedure has the advantage of being, although iterative, more analytical.

# Bibliography


[1] A. H. Reeves, French patent 852,185, October 1938, (Invention patent for PCM). Assigned to ITT.

[2] A. Reeves, "The past, present, and future of Pulse-Coded Modulation," Available from http://www.quantium.plus.com/ahr/, 1964.

[3] S. Verdú, "Fifty years of Shannon theory," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2057–2078, October 1998.

[4] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[5] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2693–2723, October 1998.

[6] T. Berger, *Rate distortion theory: a mathematical basis for data compression*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

[7] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. Urbana: Univ. of Illinois Press, 1949.

[8] C. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRNE Nat. Conv. Rec., Pt 4*, pp. 12–163, 1959.

[9] R. G. Gallager, *Information theory and reliable communication*. New York: Wiley, 1968.

[10] R. M. Gray, "Information rates of autorregressive processes," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 412–421, July 1970.

[11] H. H. Tan and K. Yao, "Evaluation of rate-distortion functions for a class of independent identically distributed sources under an absolute-magnitude criterion," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 1, pp. 59–64, January 1975.

[12] A. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IRE Trans. Inf Theory*, vol. IT-2, pp. 102–108, December 1956.

[13] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Process. Mag.*, pp. 23–50, November 1998.

[14] G. Gibson, T. Berger, T. Lookabaugh, D. Lindebergh, and R. Baker, *Digital Compression for Multimedia: Principles and Standards*. San Francisco: Morgan Kaufmann, 1998.

[15] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate-distortion function by prediction." *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3354–3364, 2008.

[16] P. Zador, "Assymptotic quantization error of continuous signals and the quantization dimension," *IEEE Trans. Inf. Theory*, vol. IT-28, pp. 239–247, March 1982.

[17] J. Bucklew and G. Wise, "Multidimensional asymptotic quantization theory with $r$th power distortion measures," *IEEE Trans. Inf. Theory*, vol. IT-28, pp. 239–247, March 1982.

[18] T. Linder and R. Zamir, "High resolution source coding for non-difference distortion measures: the rate-distortion function," *IEEE Trans. Inf. Theory*, vol. 45, pp. 533–547, 1999.

[19] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1939–1952, November 1994.

[20] N. Jayant, J. Johnston, and R. Safrenek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, no. 10, pp. 1385–1422, October 1993.

[21] J. Hawkins and S. Stevens, "The masking of pure tones and of speech by white noise," *J. Acoust. Soc. Am.*, vol. 22, pp. 6–13, 1950.

[22] W. Egan, W. Lindner, and D. McFadden, "Masking-level differences and the form of the psychometric function," *Perception & Psychoacoustics*, vol. 6, pp. 209–215, 1969.

[23] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool, 2006.

[24] S. O. Aase and T. A. Ramstad, "On the optimality of nonunitary filter banks in subband coders," *IEEE Trans. Image Process.*, vol. 4, no. 12, pp. 1585–1591, December 1995.

[25] I. Djokovic and P. Vaidyanathan, "On optimal analysis/synthesis filters for coding gain maximization," *IEEE Trans. Signal Process.*, vol. 44, no. 5, pp. 1276–1279, May 1996.

[26] P. Vaidyanathan, "Theory of optimal orthonormal subband coders," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1528–1543, June 1998.

[27] M. Nadenau, J. Reichel, and M. Kunt, "Wavelet-based color image compression: exploiting the contrast sensitivity function," *IEEE Trans. Image Process.*, vol. 12, no. 1, pp. 58–70, January 2003.

[28] A. Makur and M. Arunkamar, "Minimization of quantization noise amplification in biorthogonal subband coders," *IEEE Trans. Circuits Syst. I*, vol. 51, no. 10, pp. 2088–2091, October 2004.

[29] S. A. Jantzi, K. W. Martin, and A. S. Sedra, "Quadrature bandpass $\Sigma\Delta$ modulation for digital radio," *IEEE J. Solid-State Circuits*, vol. 32, no. 12, pp. 1935–1950, December 1997.

[30] E. King, A. Eshraghi, I. Galton, and T. Fiez, "A Nyquist-rate delta-sigma A/D converter," *IEEE J. Solid-State Circuits*, vol. 33, no. 1, pp. 45–52, January 1998.

[31] A. Eshraghi and T. Fiez, "A comparative analysis of parallel delta-sigma ADC architectures," *IEEE Trans. Circuits Syst. I*, vol. 51, pp. 450–458, 2004.

[32] G. N. Nair, F. Fagnani, S. Zampieri, and R. J. Evans, "Feedback control under data rate constraints: an overview," *Proc. IEEE*, vol. 95, no. 1, pp. 108–137, January 2007.

[33] N. Thao and M. Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Trans. Signal Process.*, vol. 42, pp. 519–531, Mar. 1994.

[34] N. T. Thao and M. Vetterli, "Reduction of the mse in $R$-times oversampled A/D conversion from $o(1/r)$ to $o(1/r^2)$," *IEEE Trans. Signal Process.*, vol. 42, pp. 200–203, 1994.

[35] ——, "Lower bound on the mean-squared error in oversampled quantization of periodic signals using vector quantization analysis," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 469–479, March 1996.

[36] V. K. Goyal, M. Vetterli, and N. T. Thao, "Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms," *IEEE Trans. Inf. Theory*, vol. 44, pp. 16–31, 1998.

[37] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[38] J. Benedetto, A. Powell, and Ö. Yilmaz, "Sigma-delta ($\Sigma\Delta$) quantization and finite frames," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 1990–2005, May 2006.

[39] J. J. Benedetto, A. M. Powell, and Ö. Yilmaz, "Second-order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions." *Appl. Comp. Harm. Analysis*, vol. 20, pp. 126–148, 2006.

[40] P. T. Boufounos and A. V. Oppenheim, "Quantization noise shaping on arbitrary frame expansions," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–12, 2006.

[41] M. Derpich, D. Quevedo, and G. Goodwin, "Conditions for optimality of scalar feedback quantization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Las Vegas, NV, 2008, pp. 3749–3752.

[42] S. Norsworthy, R. Schreier, and G. Temes (Eds.), *Delta–Sigma Data Converters: Theory, Design and Simulation*. Piscataway, NJ: IEEE Press, 1997.

[43] R. Schreier and G. Temes, *Understanding Delta-Sigma data converters*. Wiley-IEEE Press, 2004.

[44] F. Maloberti, *Data ConvertersJayant*. Dordrecht, The Netherlands: Springer, 2007.

[45] J. C. Candy and G. C. Temes, Eds., *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation*. New York: IEEE Press., 1992.

[46] R. A. Wannamaker, "Psycho-acoustically optimal noise shaping," *J. Audio Eng. Soc.*, vol. 40, no. 7/8, pp. 611–620, July/Aug. 1992.

[47] C. Dunn and M. Sandler, "Psychoacoustically optimal Sigma–Delta modulation," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 212–223, Apr. 1997.

[48] T.-C. Chang and J. P. Allebach, "Quantization of accumulated diffused errors in error diffusion," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 1960–1976, December 2005.

[49] D. Anastassiou, "Error diffusion coding for A/D conversion," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 1175–1186, 1989.

[50] P. W. Wong, "Error diffusion with delayed decision," in *SIPE/IS&T Symp. Electronic Imaging*, San Jose, CA, February 1995.

[51] H. Kato, "Trellis noise-shaping converters and 1-bit digital audio," in *112th Convention of the Audio Eng. Soc.*, 2002.

[52] P. Harpe, D. Reefman, and E. Janssen, "Efficient trellis-type Sigma Delta modulator," in *14th Convention of the Audio Eng. Soc., Paper 5845*, March 2003.

[53] D. E. Quevedo, G. C. Goodwin, and H. Bölcskei, "Multi-step optimal quantization in oversampled filter banks," in *Proc. IEEE Conf. Decis. Contr.*, 2004.

[54] D. E. Quevedo and G. C. Goodwin, "Multistep optimal analog-to-digital conversion," *IEEE Trans. Circuits Syst. I*, vol. 52, Issue 3, pp. 503–515, March 2005.

[55] N. Jayant and P. Noll, *Digital Coding of Waveforms. Principles and Approaches to Speech and Video*. Englewood Cliffs, NJ: Prentice Hall, 1984.

[56] S. K. Tewksbury and R. W. Hallock, "Oversampled, linear predictive and noise-shaping coders of order $N > 1$," *IEEE Trans. Circuits Syst.*, vol. 25, no. 7, pp. 436–447, July 1978.

[57] A.-M. Phoong and Y.-P. Lin, "Prediction-based lower triangular transform," *IEEE Trans. Signal Process.*, vol. 48, pp. 1947–1955, 2000.

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

[59] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, New Jersey: Prentice-Hall, 1993.

[60] N. Fliege, *Multirate Digital Signal Processing: Multirate Systems, Filter Banks, Wavelets*. New York, NY: John Wiley & Sons, 1994.

[61] H. Bölcskei and F. Hlawatsch, "Noise reduction in oversampled filter banks using predictive quantization," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 155–172, Jan 2001.

[62] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 2001.

[63] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, N.J: Wiley-Interscience, 2006.

[64] P. Noll and R. Zelinski, "Bounds on quantizer performance in the low bit-rate region," *IEEE Trans. Commun.*, vol. COM-26, no. 2, pp. 300–304, February 1978.

[65] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 5, pp. 676–683, September 1968.

[66] E. Janssen and D. Reefman, "Super-audio CD: an introduction," *IEEE Signal Processing Magazine*, pp. 83–90, July 2003.

[67] W. R. Bennet, "Spectrum of quantized signals," *Bell Syst. Tech J.*, vol. 27, pp. 446–472, July 1948.

[68] J. Nilsson, "Real-time control systems with delays," Ph.D. dissertation, Lund Institute of Technology, 1998.

[69] G. C. Goodwin, S. F. Graebe, and M. E. Salgado, *Control System Design*. Prentice-Hall, 2001.

[70] T. Linder and R. Zamir, "Causal coding of stationary sources and individual sequences with high resolution," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 662–680, February 2006.

[71] A. Gersho, "Principles of quantization," *IEEE Trans. Circuits Syst.*, vol. 25, no. 7, pp. 427– 436, July 1978.

[72] W. R. Gardner and B. D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Trans. Speech, and Audio Processing*, vol. 3, no. 5, pp. 367–381, September 1995.

[73] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. Heath, "Design of linear equalizers optimized for the structural similarity index." *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 857–872, June 2008.

[74] R. A. McDonald and P. M. Schultheiss, "Information rates of Gaussian signals under criteria constraining the error spectrum." *Proc. IEEE*, vol. 52, no. 4, pp. 415–416, April 1964.

[75] R. Vafin and W. B. Kleijn, "Rate-distortion optimized quantization in multistage audio coding," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 1, pp. 311–320, January 2006.

[76] R. W. Floyd and L. Steinberg, "An adaptive algorithm for spatial gray scale," in *Proc. SID Int. Symp. Dig. Tech. Papers*, 1976, pp. 36–37.

[77] F. Baqai, J.-H. Lee, A. Agar, and J. Allebach, "Digital color halftoning," *Signal Processing Magazine, IEEE*, vol. 22, no. 1, pp. 87–96, Jan 2005.

[78] S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds., *Delta–Sigma Data Converters: Theory, Design and Simulation*. Piscataway, N.J.: IEEE Press, 1997.

[79] H. Spang, III and P. Schultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Comm. Syst.*, vol. CS-10, no. 4, pp. 373–380, Dec. 1962.

[80] R. Brainard and J. Candy, "Direct-feedback coders: design and performance with television signals." *Proc. IEEE*, vol. 57, no. 7, pp. 776–786, July 1969.

[81] P. Noll, "On predictive quantizing schemes," *Bell. Syst. Tech. J.*, vol. 57, no. 5, pp. 1499–1532, May-June 1978.

[82] B. S. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 3, pp. 247–254, June 1979.

[83] D. Stacey, R. Frost, and G. Ware, "Error spectrum shaping quantizers with non-ideal reconstruction filters and saturating quantizers," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 3, 1991, pp. 1905–1908.

[84] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, May 1993.

[85] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A theory of non-subtractive dither," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 499–516, Feb. 2000.

[86] P. Moulin, M. Anitescu, and K. Ramchandran, "Theory of rate-distortion-optimal, constrained filterbanks– application to IIR and FIR biorthogonal designs," *IEEE Trans. Signal Process.*, vol. 48, no. 4, pp. 1120–1132, April 2000.

[87] J. Tuqan and P. P. Vaidyanathan, "Statistically optimum pre- and postfiltering in quantization." *IEEE Trans. Circuits Syst. II*, vol. 44, no. 1, pp. 1015–1031, December 1997.

[88] M. Gerzon and P. G. Craven, "Optimal noise shaping and dither of digital signals," in *87th Convention of the AES, New York, NY, preprint 2822*, Oct. 1989.

[89] E. Kimme and F. Kuo, "Synthesis of optimal filters for a feedback quantization system." *IEEE Trans. Circuit Theory*, vol. CT-10, pp. 405–413, September 1963.

[90] O. Guleryuz and M. Orchard, "On the DPCM conpression of Gaussian autoregressive sequences," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 945–956, March 2001.

[91] Cvetković, "Resilience properties of redundant expansions under additive noise and quantization," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 644–656, Mar. 2003.

[92] H. Bölcskei, "Noise shaping quantizers of order $L > 1$ for "general" frame expansions." presented at the Workshop on Coarsely Quantized Redundant Representations of Signals, Banff , Alberta, Canada, 2006.

[93] M. Lammers, A. M. Powell, and Ö. Yilmaz, "Alternative dual frames for digital-to-analog conversion in sigma-delta quantization," *Adv. Comput. Math.*, July 2008.

[94] D. Hui and D. L. Neuhoff, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization." *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 957–977, March 2001.

[95] I. Daubechies and R. DeVore, "Approximating a bandlimited function using very coarsely quantized data: A family of stable Sigma-Delta modulators of arbitrary order," *Ann. of Math.*, vol. 158, no. 2, pp. 679–710, 2003.

[96] C. S. Güntürk, "One-bit Sigma-Delta quantization with exponential accuracy," *Commun. Pure Appl. Math.*, vol. 56, no. 11, pp. 1608–1630, 2003.

[97] Z. Cvetković and M. Vetterli, "Error-rate characteristics of oversampled analog-to-digital conversion," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1961–1964, 1998.

[98] Z. Cvetković and I. Daubechies, "Single-bit oversampled A/D compression with exponential accuracy in the bit-rate," in *Proc. Data Comp. Conf.*, Mar 2000.

[99] Z. Cvetković and M. Vetterli, "On simple oversampled A/D conversion in $L^2(\mathbb{R})$," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 146–154, January 2001.

[100] P. Vaidyanathan, "Theory and design of $M$-channel maximally decimated quadrature mirror filters with arbitrary $M$, having the perfect-reconstruction property," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 4, pp. 476–492, April 1987.

[101] R. D. Koilpillai and P. P. Vaidyanathan, "Cosine-modulated FIR filter banks satisfying perfect reconstruction," *IEEE Trans. Signal Process.*, vol. 40, pp. 770–783, Apr. 1992.

[102] Y.-P. Lin and P. P. Vaidyanathan, "Linear phase cosine modulated maximally decimated filter banks with perfect reconstruction," *IEEE Trans. Signal Process.*, vol. 42, pp. 2525–2539, Nov. 1995.

[103] R. A. Haddad and K. Park, "Modeling, analysis, and optimum design of quantized $M$-band filter banks," *IEEE Trans. Signal Process.*, vol. 43, no. 11, pp. 2540–2579, November 1995.

[104] Z. Cvetković and M. Vetterli, "Oversampled filter banks," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1245–1255, May 1998.

[105] H. Bölcskei and F. Hlawatsch, "Oversampled cosine modulated filter banks with perfect reconstruction," *IEEE Trans. Circuits Syst. II*, vol. 45, no. 8, pp. 1057–1071, Aug. 1998.

[106] P. H. Westerink and J. Biemond, "Scalar quantization error analysis for image subband coding using QMF's," *IEEE Trans. Signal Process.*, vol. 40, no. 2, pp. 421–428, February 1992.

[107] J. Kovačević, "Subband coding systems incorporating quantizer models," *IEEE Trans. Image Process.*, vol. 4, no. 5, pp. 543–553, May 1995.

[108] A. Dembo and D. Malah, "Statistical design of analysis/synthesis systems with quantization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 3, pp. 328–341, March 1988.

[109] P. Moulin, "A relaxation algorithm for minimizing the $L^2$ reconstruction error in 2-D nonorthogonal subband coding," in *1st IEEE Int. Conf. Image Processing*, Austin, TX, 1994, pp. 908–912.

[110] ——, "A multiscale relaxation algorithm for SNR maximization in nonorthogonal subband coding," *IEEE Trans. Image Process.*, vol. 4, no. 9, pp. 1269–1281, September 1995.

[111] K. C. Aas and C. T. Mullis, "Minimum mean-squared error transform coding and subband coding," *IEEE Trans. Inf. Theory*, vol. 42, no. 4, pp. 1179–1192, July 1996.

[112] M. G. Strintzis, "Optimal subband coders of quantized multidimensional signals," *IEEE Trans. Circuits Syst. II*, vol. 47, no. 8, pp. 757–770, August 2000.

[113] M. K. Mıhçak, P. Moulin, M. Anitescu, and K. Ramchandran, "Rate-distortion-optimal subband coding without perfect-reconstruction constraints," *IEEE Trans. Signal Process.*, vol. 49, no. 3, pp. 542–557, March 2001.

[114] P. Vaidyanathan and T. Chen, "Statistically optimal synthesis banks for subband coders," in *Asilomar Conf. Signals, Syst. Comput.*, vol. 2, 1994, pp. 986–990.

[115] J. W. Woods and S. D. O'Neil, "Subband coding of images," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 5, pp. 1278–1288, October 1986.

[116] T. R. Fischer, "On the rate-distortion efficiency of subband coding," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 426–428, March 1992.

[117] P. W. Wong, "Rate distortion efficiency of subband coding with crossband prediction," *IEEE Trans. Inf. Theory*, vol. 43, pp. 352–356, 1997.

[118] M. S. Derpich, E. I. Silva, D. E. Quevedo, and G. C. Goodwin, "On optimal perfect reconstruction feedback quantizers," *IEEE Trans. Signal Process.*, vol. 56, no. 8, Part 2, pp. 3871–3890, August 2008.

[119] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York, NY: Academic Press, 1970.

[120] A. Habibi and R. Hershel, "A unified representation of differential pulse-coded modulation (DPCM) and transform coding systems," *IEEE Trans. Commun.*, vol. 22, no. 5, pp. 692–696, May 1974.

[121] V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Process. Mag.*, vol. 18, pp. 9–21, Sept. 2001.

[122] T. Ericson, "A result on delay-less information transmission," Int. Symp. Inform. Theory, Grignano, Italy, June 1979.

[123] N. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems," Int. Symp. Inform. Theory, Grignano, Italy, June 1979.

[124] ——, "On optimal finite-state digital transmission systems," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 167–186, March 1982.

[125] D. Neuhoff and R. Gilbert, "Causal source codes," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 5, pp. 701–713, September 1982.

[126] R. Zamir and M. Feder, "On universal quantization by randomized uniform/latice quantizers," *IEEE Trans. Inf. Theory*, vol. 38, pp. 428–436, 1992.

[127] M. S. Derpich, J. Østergaard, and G. C. Goodwin, "The quadratic Gaussian rate-distortion function for source uncorrelated distortions," in *Proc. Data Compression Conf.*, Snowbird, UT, March 2008, pp. 73–82.

[128] D. Luenberger, *Optimization by Vector Space Methods*. London: John Wiley and Sons, Inc., 1969.

[129] R. M. Gray, "Toeplitz and circulant matrices: a review," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[130] D. Marco and D. L. Neuhoff, "The validity of the additive noise model for uniform scalar quantizers," *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1739–1755, May 2005.

[131] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. Commun.*, vol. 12, pp. 162–165, Dec. 1964.

[132] J. Ziv, "On universal quantization," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 3, pp. 344–347, May 1985.

[133] S. P. Lipschitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, no. 5, pp. 355–375, May 1992.

[134] R. A. Wannamaker, "The theory of dithered quantization," Ph.D. dissertation, Dept. of Applied Math., Univ. of Waterloo, Waterloo, ON, Canada, 1997.

[135] C. H. Bae, J. H. Ryu, and K. W. Lee, "Suppression of harmonic spikes in switching converter output using dithered Sigma–Delta modulation," *IEEE Trans. Ind. Appl.*, vol. 38, no. 1, pp. 159–166, Jan./Feb. 2002.

[136] E. I. Silva, G. C. Goodwin, D. E. Quevedo, and M. S. Derpich, "Optimal noise shaping for networked control systems," in *Proc. Europ. Contr. Conf.*, Kos, Greece, July 2007.

[137] J. G. Kenney and L. R. Carley, "CLANS: a high-level synthesis tool for high resolution data converters," in *IEEE Conf. Computer-Aided Design*, November 1988, pp. 496–499.

[138] R. Eschbach, Z. Fan, K. T. Knox, and G. Marcu, "Threshold modulation and stability in error diffusion," *IEEE Signal Process. Mag.*, pp. 39–50, July 2003.

[139] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*. Englewood Cliffs, New Jersey: Prentice Hall Inc., 1963.

[140] R. Weinstock, *Calculus of Variations with Applications to Physics and Engineering*. New York: Dover Publications, Inc., 1974.

[141] M. S. Derpich, E. I. Silva, D. E. Quevedo, and G. C. Goodwin, "Optimal noise-shaping DPCM," available from http://msderpich.no-ip.org.

[142] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate-distortion function by prediction," in *Int. Symp. Information Theory*, Seattle, USA, July 2006, pp. 803–807.

[143] G. Stein, "Respect the unstable," *IEEE Control System Magazine*, vol. 23, no. 4, August 2003.

[144] G. F. Carrier, M. Krook, and C. Pearson, *Functions of a Complex Variable: Theory and Technique*. Ithaca, N.Y.: Hod Books, 1983.

[145] M. M. Serón, J. H. Braslavsky, and G. C. Goodwin, *Fundamental Limitations in Filtering and Control*. Springer-Verlag, London, 1997.

[146] R. L. Burden and J. D. Faires, *Numerical Analysis*, ser. The Prindle, Weber & Schmidt series in mathematics. Boston: PWS-Kent Pub. Co., 1993.

[147] S. Boyd and L. Vandenberghe, *Convex Optimization*. Springer, 2004.

[148] J. O'Neal Jr., "Bounds on subjective performance measures for source encoding systems," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 3, pp. 224–231, May 1971.

[149] H. J. Godwin, *Inequalities on Distribution Functions*, 1st ed., S. M.G. Kendall, M.A., Ed. London: Charles Griffin & Company Limited, 1964.

[150] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*. Cambridge Univ. Press, 1959.

[151] Y. Yamada, S. Tazaki, and R. M. Gray, "Asymptotic performance of block quantizers with difference distortion measures," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 1, pp. 6–14, January 1980.

[152] J. Li, N. Chaddha, and R. M. Gray, "Asymptotic performance of vector quantizers with a perceptual distortion measure," *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1082–1091, May 1999.

[153] E. Martinian, G. W. Wornell, and R. Zamir, "Source coding with distortion side information," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4638–4665, October 2008.

[154] B. Friestedt and L. Gray, *A Modern Approach to Probability Theory*, ser. Probability and its applications. Boston: Birkhäuser, 1997.

[155] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covariance constraint," *IEEE Trans. Inf. Theory*, vol. 47, pp. 3072–3081, 2001.

[156] L. G. Roberts, "Picture coding using pseudo-random noise," *IRE Trans. Inf. Theory*, vol. IT–8, pp. 145–154, February 1962.

[157] B. Lippel and M. Kurland, "The effect of dither on luminance quantization of pictures," *IEEE Trans. Commun. Technol.*, vol. COM–19, no. 6, pp. 879–888, December 1971.

[158] N. S. Jayant and L. R. Rabiner, "The application of dither to the quantization of speech signals," *The Bell Syst. Tech. J.*, vol. 51, no. 6, pp. 1293–1304, 1972.

[159] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, UK: Cambridge University Press, 1985.

[160] M. Derpich, J. Østergaard, and D. Quevedo, "Achieving the quadratic Gaussian rate-distortion function for source uncorrelated distortions," (available at http://arxiv.org/abs/0801.1718v3 ).

[161] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Inf. Theory*, vol. 42, no. 4, pp. 1152–1159, July 1996.

[162] J. Massey, "Causality, feedback and directed information," Available from http://citeseer.ist.psu.edu/massey90causality.html, 1990.

[163] M. Marcus, "An eigenvalue inequality for the product of normal matrices," *Amer. Math. Monthly*, vol. 63, no. 3, pp. 173–174, March 1956.

[164] G. Kramer, "Directed information for channels with feedback." Ph.D. dissertation, Swiss federal institute of technology, 1998.

[165] D. L. Mary and D. T. M. Slock, "A theoretical high-rate analysis of causal verus unitary online transform coding," *IEEE Trans. Signal Process.*, vol. 4, no. 4, pp. 1472–1482, April 2006.

[166] J. J. Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun. Syst.*, vol. COM-11, pp. 289 – 296, September 1963.

[167] S. Rao and W. Pearlman, "Analysis of linear prediction, coding and spectral estimation from subbands," *IEEE Trans. Inf. Theory*, vol. 42, no. 4, pp. 1160–1178, July 1996.

[168] T. A. Ramstad, S. O. Aase, and J. H. Husøy, *Subband Compression of Images: Principles and Examples*, ser. Advances in image communication. Amsterdam, The Netherlands: Elsevier, 1995.

[169] M. Pisnker and A. Gorbunov, "Epsilon-entropy with delay for small mean-square reproduction error," *Probl. Inf. Transm.*, vol. 23, pp. 91–95, 1987, translation from Problemi Peredachi Informatsii, vol. 23, no. 2, pp. 3–8, April-June 1987.

[170] A. Gorbunov and M. Pinsker, "Asymptotic behavior of nonanticipative epsilon-entropy for Gaussian processes," *Probl. Inf. Transm.*, vol. 27, no. 4, pp. 361–365, 1991, translation from Problemi Peredachi Informatsii, vol. 27, no. 4, pp. 100–104, October-December 1991.

[171] P. Ishwar and K. Ramchandran, "On decoder-latency versus performance tradeoffs in differential predictive coding," in *Proc. Int. Conf. Image Proc.*, 2004, pp. 1097–1100.

[172] B. Agrawal and K. Shenoi, "Design methodology for $\Sigma\Delta$M," *IEEE Trans. Commun.*, vol. COM-31, pp. 360–370, 1983.