Novel Schemes for Sigma-Delta Modulation: From Improved Exponential Accuracy to Low-Complexity Design

by

Felix Krahmer

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics

New York University

September 2009

Professor Percy Deift

Professor Sinan Güntürk

To my family.

Acknowledgements

First, I would like to thank my advisors Percy Deift and Sinan Güntürk for all their time and support, for our frequent meetings and fruitful discussions. I always appreciated their willingness and excitement to advise me jointly, to share their different perspectives on a problem that is between their research areas, and to offer their advice and help on issues that often reached far beyond the mathematical content. I am indebted to both of them.

My thesis would not have been possible in this form without Nguyen T. Thao. It was through the discussions with him that I was first exposed to the engineers' perspective on $\Sigma\Delta$ modulation, which was crucial for relating my results to the application. My recent interaction with Yonina Eldar also helped shaping my understanding of the engineers' point of view on the subject.

It was through some of Götz Pfander's lectures during my undergraduate studies at International University Bremen that I became interested in Signal Processing. I thank him for his mentorship during our joint research project and beyond. The summer academy he co-organized in Bremen in 2007 has offered me insight into many areas in engineering and fueled my interest in collaborating directly with engineers on practically relevant problems.

I thank Jinho Baik, Chee Yap, Gitta Kutyniok, Holger Rauhut, and Rachel Ward for our collaborations and scientific interactions.

A special thanks goes to my friend and collaborator Michael Burr, who has tremendously helped me improve the writing of this thesis. Fedor Soloviev has been my office mate and my close friend for the whole time at the Courant Institute, I thank him for all the time we spent together. I also thank Kela, Oren, Abhishek, Hans S., Hans K., Juliana, Nam and many others at the Courant Institute and at

NYU for their friendship, support, and our common activities.

I was supported through various sources throughout my study. I thank the Graduate School of Arts and Sciences at NYU for the Henry McCracken fellowship and for the GSAS Dean's travel grant, the Courant Institute for additional summer support and the Charles M. Newman Fellowship, and Cathleen S. Morawetz for the Morawetz Fellowship.

I thank the NYU Jazz Choir, the German St Paul's Church, as well as Imke, Karsten, and André for providing a balance through non-academic activities.

Finally, I thank my girlfriend Anne, my parents, my grandmother and my brother Constantin for all the support they provided over the distance and during their visits.

Abstract

The central theme of this thesis is one-bit quantization of bandlimited signals via Sigma-Delta modulation. In this commonly used analog-to-digital conversion method, the signal of interest is represented by a ± 1-valued sequence that is computed recursively from regular samples of the signal via a difference equation. The key feature of the method is that the low frequency content of the quantized representation approximates the signal: The larger the oversampling rate λ with respect to the Nyquist frequency, the higher the accuracy of the reconstruction that is achievable.

It is known that exponential accuracy with an error decay rate $O(2^{-r\lambda})$ for some rate constant $r > 0$ is achievable via Sigma-Delta modulation with modulators of increasing order. In this thesis, we first construct a family of schemes which gives a better rate constant r than is known for oversampled one-bit quantization. The construction builds on an idea by Güntürk and proceeds by solving an optimization problem posed in his work. En route, the solution establishes a connection between Sigma-Delta modulation and the theory of orthogonal polynomials.

Second, we prove stability results for Sigma-Delta modulators involving recursion filters with rational transfer functions; stability is crucial to achieve satisfactory approximation. Such modulators are commonly used in practice because the associated analog circuits are of low complexity. Nevertheless, prior to this thesis, a rigorous stability analysis for such modulators was not available. We construct the first family of provably stable modulators of this type for all orders. Also, we introduce a novel, generalized stability criterion for Sigma-Delta modulation.

Contents

Dedication iii

Acknowledgements iv

Abstract vi

List of Figures ix

Introduction 1

1 Two views on one-bit quantization **5**

1.1 The engineering perspective 5

1.1.1 Sampling and reconstruction 6

1.1.2 Filters in discrete signal processing 11

1.1.3 Quantization 14

1.1.4 Noise shaping 15

1.2 The mathematical perspective 19

1.2.1 General setup 19

1.2.2 Error decay for m-th order $\Sigma\Delta$ modulators 22

1.2.3 Superpolynomial error decay 32

1.2.4 A basic stability criterion for greedy quantization 33

2 Optimizing minimally supported filters **36**

2.1 An optimization problem for filters with minimal support 36
2.2 The relaxed minimization problem for optimal filters 39
2.3 Some useful properties of Chebyshev polynomials 44
2.4 Solution of the relaxed minimization problem 47
2.5 Asymptotics for the relaxed and the discrete minimization problem 57

3 Stability analysis for MCR ΣΔ modulators 68

3.1 MCR modulators 68
3.2 A criterion for the roots of the denominator 70
3.3 Application to schemes with linear quantization rules 80

4 A novel stability criterion 81

4.1 A stability criterion for approximately greedy quantization rules . . 81
4.2 General formulation of the stability criterion 84
4.3 Application to MCR ΣΔ modulators 86

Bibliography 88

List of Figures

1.1 Block diagram of a $\Sigma\Delta$ modulator 17

2.1 The minimizer $\mathbf{x}$ of the relaxed problem for $m = 18$ and $\gamma = 1.5$ and the corresponding minimal subordinate integer sequence $\mathbf{n}$ . . . 65

Introduction

In this thesis we study the design and analysis of analog-to-digital (A/D) conversion schemes for audio signals. An A/D conversion scheme represents a given input signal by a finite number of symbols chosen from a given finite alphabet. This digital representation should be designed to capture the information of the signal to any given desired degree of accuracy [14]. A/D conversion is of great importance in modern signal processing, as a digital signal representation has numerous advantages compared to the original analog representation. First, analog signal processing and communications only perform an approximation of the targeted mathematical operations (including analog distortions and added noise), while digital signal processing and communication perform exact operations. In wireless communication, for example, techniques like orthogonal frequency-division multiplexing (OFDM) allow for different senders to transmit independent digital signals over the same channel without data losses caused by interferences between the different signals. The idea is that different senders use different frequency bands; working with digital signals ensures that each signal remains in the assigned band. Furthermore, many error prevention and error correction techniques based on results in coding theory only apply to digital signals. Second, digital signals allow for easier and more accurate storage, as they can be equivalently represented by a finite sequence of *bits*, i.e., a sequence that only assumes the values $\{0, 1\}$. Storing bits is easy to implement, as only two possible values have to be distinguished – in contrast to a continuum for analog signals. Now, while digital operations are exact in the sense that they are (almost) exactly reproducible, the main source of error lies in the discretization of the signals to be processed. For this reason, efficient and accurate techniques for A/D and D/A conversion are of great importance.

For the purposes of A/D-conversion, audio signals are usually modeled as bounded bandlimited functions. For such functions, the well-known Shannon-Nyquist sampling theorem applies, viz. a bandlimited function f can be reconstructed exactly from its sample values $y_n = f\left(\frac{n}{\lambda}\right)$ as long as the sampling frequency λ lies above the critical (Nyquist) rate λ_0. At the level of circuit implementation, reconstruction is realized by a low-pass filter. A good digital representation is a sequence (q_n) of *quantized values* chosen from a finite set such that the same low-pass filter yields a good approximation to f when applied to the q_n's [14].

In *pulse code modulation* (PCM), the signal is sampled at a fixed frequency $\lambda \approx \lambda_0$, and q_n corresponds to a truncated binary representation of x_n. For better accuracy, one increases the precision of the approximation of each sample value. In *oversampled coarse quantization*, the set of admissible values for q_n is fixed and higher accuracy is achieved by increasing λ. A common special case occurs in *one-bit quantization* schemes, which work with the admissible set $\{-1, 1\}$. Oversampled coarse quantization is possible because the redundancy of the sequence of sample values y_n increases as λ is increased. However, it is still nontrivial to design a procedure to find a sequence (q_n) that guarantees accurate approximation when $\lambda \to \infty$.

From the viewpoint of circuit engineering, oversampled coarse quantization means low-cost analog hardware because increasing the sampling rate is cheaper than refining the quantization. For this reason, oversampling data converters, in particular, Sigma-Delta ($\Sigma\Delta$) modulators, have become more popular than Nyquist-rate converters for low to medium-bandwidth signal classes, such as audio signals [13]. Further advantages of oversampled coarse quantization include a more even distribution of the bit significance [3]: In the binary representation of

a real number, the first digits have higher significance than later digits. Thus bit errors occurring in the first few digits of a quantized value in a PCM scheme result in grave reconstruction errors, whereas bit errors in other digits do not. Many results from coding theory are based on the assumption of a bit stream with equal significance and hence do not apply here. In the one-bit quantization scheme, on the other hand, the individual bits carry equal significance. In this case, bit errors always have the same effect on the reconstruction error and coding theory techniques apply directly.

This thesis is concerned with the approximation theory of oversampled coarse quantization, in particular, one-bit quantization of bandlimited functions. Oversampled coarse quantization has frequently been discussed in the engineering literature (e.g., see [13]), and recently there have been a series of more mathematically oriented papers on the subject (e.g., [4, 7, 8, 9, 1]).

In Chapter 1 of this thesis, we discuss both the engineering perspective and the mathematical perspective on one-bit quantization: In Section 1.1, we motivate the core concepts of digital signal processing from the engineering point of view without dwelling on the precise mathematical formulations and use these ideas to motivate $\Sigma\Delta$ modulation; in Section 1.2, we then embed these ideas into a more precise mathematical framework.

In Chapter 2, we focus on the reconstruction error that arises in $\Sigma\Delta$ modulation. Our analysis is based on the idea, first introduced in [4], to optimize the bounds on the error decay by choosing different circuit architectures for different sampling rates λ. This technique has also been employed in [7] to show that exponential accuracy in the oversampling ratio λ can be achieved by appropriate one-bit $\Sigma\Delta$ modulation schemes. Prior to this thesis, the best achievable error

decay rate for these schemes was $O(2^{-r\lambda})$ with $r \approx 0.076$. Chapter 2 improves the best achievable rate further to $r \approx 0.102$. It is known that any 1-bit quantization scheme has to obey $r < 1$ [3, 7] (However, it is not known if this upper bound is tight). Our method employed in this chapter draws from the theory of orthogonal polynomials; in particular, it relates the filters used in our optimized construction to the zero sets of Chebyshev polynomials of the second kind. For the convenience of the reader, in Section 2.3 we collect some properties and identities for Chebyshev polynomials which are used in the proofs of our results.

The $\Sigma\Delta$ modulators considered in Chapter 2 employ finite filters with polynomial transfer functions. In practice, it is often preferred to employ $\Sigma\Delta$ modulators with rational transfer functions such that the associated analog circuits are of minimal complexity. However, little was known about the rigorous error analysis for this class of modulators. In Chapter 3, we provide such an error analysis showing that superpolynomial error decay can be achieved using modulators in this class.

The results in Chapters 2 and 3 are based on a well-known stability criterion, which works only for schemes that employ a particular, so-called greedy, quantization rule. In Chapter 4 we extend this stability criterion to apply to a more general class of quantization rules. The resulting generalized criterion is then used to make generalized inferences for schemes that employ the greedy rule.

Chapter 1

Two views on one-bit quantization

1.1 The engineering perspective

In this section we motivate the mathematical constructs examined in the following chapters from the engineering perspective. We follow the standard texts *Signals and Systems* by Oppenheim, Willsky and Nawab [15] and *Discrete-Time Signal Processing* by Oppenheim, Schafer and Buck [14].

In this section we deliberately refrain from discussing the underlying spaces of functions or distributions and the specific properties of the mathematical operations considered. Also we do not address problems which may arise in certain cases from lacking smoothness or decay properties of the functions involved. In Section 1.2, we provide a precise mathematical formulation of the underlying concepts.

The audio signals to be quantized are modeled as bounded Ω-*bandlimited* functions f, i.e., the Fourier transform of the signal $\widehat{f}(\xi) = \int_{-\infty}^{\infty} f(t)e^{-2\pi it}dt$ is supported

in a given interval $\left[-\frac{\Omega}{2}, \frac{\Omega}{2}\right]$, and the amplitude of the function is less than some constant μ. We consider the function and its Fourier transform as two different representations of the same signal, where the variable t in the representation $f(t)$ is referred to as the *time* and the variable ξ in the representation $\widehat{f}(\xi)$ is referred to as the *frequency*.

An example of a Ω-bandlimited function for $\Omega = 1$ is the function with the time representation

$$f^{(ex)}(t) = \frac{3}{4} + \frac{1}{4}\sin(3x). \tag{1.1}$$

Indeed, the corresponding frequency representation

$$\widehat{f^{(ex)}}(\xi) = \frac{3}{4}\delta^{(0)}(\xi) + \frac{1}{8i}\left[\delta^{\left(\frac{3}{2\pi}\right)}(\xi) - \delta^{\left(-\frac{3}{2\pi}\right)}(\xi)\right], \tag{1.2}$$

where $\delta^{(b)}$ denotes the Dirac delta function centered at b, is supported in $\left[-\frac{1}{2}, \frac{1}{2}\right]$. This bandlimited function shall serve as an example to illustrate the following concepts.

1.1.1 Sampling and reconstruction

Sampling lies at the core of all digital signal processing. The goal is to reduce a continuous signal f to a discrete representation. Most commonly, the signal is represented by its instantaneous values $f(t_n)$ at a discrete sequence of time instances t_n, $n \in \mathbb{Z}$. We refer to the values $f(t_n)$ as the *sampled values* or just *samples* of the signal. In this thesis, we focus on *uniform sampling at rate* λ, where the time instances are chosen to be $t_n = \frac{n}{\lambda}$.

For example, sampling the signal $f^{(ex)}$ introduced above uniformly at rate $\lambda =$

1.5 yields the samples

$$\ldots, \frac{3}{4} + \frac{1}{4}\sin(-4), \frac{3}{4} + \frac{1}{4}\sin(-2), \frac{3}{4}, \frac{3}{4} + \frac{1}{4}\sin(2), \frac{3}{4} + \frac{1}{4}\sin(4), \ldots \tag{1.3}$$

When listing the samples of a signal like in Equation (1.3), it is always understood that it is also known which samples correspond to which time instances.

In signal processing, sampling a signal f is modeled as modulating $f(t)$ with a sum of unit impulses centered at each of the sampling time instances. The advantage of such a representation is that the resulting distribution embeds better in the formalism of analog signal processing than the sequence of sampled values: As we will see, for example, the D/A converter can be interpreted as an analog low-pass filter.

In such a representation, the summand corresponding to t_i is obtained by multiplying the function $f(t)$ by the Dirac delta function $\delta^{(t_i)}$. Accordingly, sampling the function f at t_0 results in the function

$$f(t)\delta^{(t_0)} = f(t_0)\delta^{(t_0)} \tag{1.4}$$

and periodic sampling at $\left\{t_n = \frac{n}{\lambda}\right\}_{n\in\mathbb{Z}}$ results in

$$f_\lambda(t) = f(t) \sum_{n=-\infty}^{\infty} \delta^{\left(\frac{n}{\lambda}\right)}(t) = \sum_{n=-\infty}^{\infty} f\left(\frac{n}{\lambda}\right) \delta^{\left(\frac{n}{\lambda}\right)}(t)\,. \tag{1.5}$$

We call f_λ the *sampled function* corresponding to a function f and the sampling rate λ. Clearly, the sampled function carries the exact same information as the sequence of sampled values with the associated time instances. The sampled function of the

signal $f^{(ex)}$ corresponding to rate $\lambda = 1.5$ is given by

$$f_{1.5}^{(ex)}(t) = \sum_{n=-\infty}^{\infty} \left(\frac{3}{4} + \frac{1}{4}\sin(2n) \right) \delta^{\left(\frac{n}{1.5}\right)}(t)\,. \tag{1.6}$$

The sampled function f_λ has the frequency representation

$$\widehat{f_\lambda}(\xi) = \widehat{f} * \left(\sum_{n=-\infty}^{\infty} \delta^{\left(\frac{n}{\lambda}\right)} \right)^{\wedge}(\xi). \tag{1.7}$$

This representation can be simplified by noting that the Fourier transform of the impulse train

$$s(t) = \sum_{n=-\infty}^{\infty} \delta^{\left(\frac{n}{\lambda}\right)}(t) \tag{1.8}$$

is given by

$$\widehat{s}(\xi) = \lambda \sum_{k=-\infty}^{\infty} \delta^{(k\lambda)}(\xi)\,. \tag{1.9}$$

This well-known fact can be proved in the context of the theory of distributions (for a justification from the engineering viewpoint see [15]). Here we will not provide a proof. Instead, Section 1.2 will introduce a precise mathematical framework that does not use delta functions; the corresponding statements will be proved there.

Using Equation (1.9), $\widehat{f_\lambda}(\xi)$ can be expressed as:

$$\widehat{f_\lambda}(\xi) = \widehat{f} * \left(\lambda \sum_{k=-\infty}^{\infty} \delta^{(k\lambda)} \right)(\xi) = \lambda \sum_{k=-\infty}^{\infty} \widehat{f}\,(\xi - k\lambda)\,. \tag{1.10}$$

Up to a constant, the frequency domain representation of the signal uniformly sampled at rate λ is a superposition of infinitely many copies $\widehat{f}$, shifted by integer

multiples of λ. For $f^{(ex)}$ and a sampling rate $\lambda = 1.5$, we obtain

$$\widehat{f_{1.5}^{(ex)}}(\xi) = 1.5 \sum_{k=-\infty}^{\infty} \frac{3}{4}\delta^{(1.5k)}(\xi) + \frac{1}{8i}\left[\delta^{\left(\frac{3}{2\pi}+1.5k\right)}(\xi) - \delta^{\left(-\frac{3}{2\pi}+1.5k\right)}(\xi)\right] \tag{1.11}$$

If an Ω-bandlimited signal is sampled at a rate $\lambda > \Omega$, the summands $\widehat{f}(\cdot - k\lambda)$ have disjoint supports. That is, the sampled function carries the complete information of the signal, and the signal can be recovered mathematically by multiplying the frequency domain representation of f_s by a function $\widehat{\varphi}$ that satisfies

$$\widehat{\varphi}(\xi) = \begin{cases} 0 \text{ for } \xi > \frac{\lambda}{2} \\ 1 \text{ for } \xi < \frac{\Omega}{2} \end{cases} \tag{1.12}$$

and dividing the result by λ to adjust the constant. In the time domain, this reconstruction method is described by the formula

$$\begin{aligned} f &= \frac{1}{\lambda} f_\lambda * \varphi. \\ &= f\left(\frac{n}{\lambda}\right) \sum_{n=-\infty}^{\infty} f\left(\frac{n}{\lambda}\right) \delta^{\left(\frac{n}{\lambda}\right)}(t) * \varphi \\ &= \frac{1}{\lambda} \sum_{n=-\infty}^{\infty} f\left(\frac{n}{\lambda}\right) \varphi\left(t - \frac{n}{\lambda}\right) \end{aligned} \tag{1.13}$$

This formula is also known as the *Shannon-Nyquist Sampling Theorem*. For example, as $f^{(ex)}$ is 1-bandlimited and $\lambda = 1.5 > \Omega = 1$, the summand corresponding to the index k in Equation (1.11) is supported in $[1.5k - 0.5, 1.5k + 0.5]$. As these intervals do not overlap, we can choose φ as in Equation (1.12) such that $\widehat{\varphi}\widehat{f_{1.5}^{(ex)}} = \widehat{f^{(ex)}}$.

As this procedure annihilates the high-frequency components of the function

while leaving the low-frequency content unchanged, it is realized in the circuit implementation by a low-pass filter (for a detailed discussion on low-pass filters see [15]). In practice, the function $\widehat{\varphi}$ corresponding to a low-pass filter will only be approximately constant on the interval $[-\frac{\Omega}{2}, \frac{\Omega}{2}]$, but for this thesis, we will assume an ideal low-pass filter that exactly satisfies Relation (1.12). Such a low-pass filter will allow exact reconstruction of a bandlimited signal from its samples.

If $\lambda < \Omega$ or if the signal is not bandlimited, then the supports of the shifted copies of $\widehat{f}$ will in general not be disjoint. Hence, the different summands of $\widehat{f_s}$ will interfere, a phenomenon referred to as *aliasing*. As a consequence, each $\widehat{f_s}$ corresponds to more than one $\widehat{f}$, and the sample cannot be uniquely recovered from its samples. For example, if $f^{(ex)}$ is sampled at rate $\lambda = \frac{3}{2\pi}$, we obtain

$$\begin{aligned}\widehat{f^{(ex)}_{\frac{3}{2\pi}}}(\xi) &= \frac{3}{2\pi}\sum_{k=-\infty}^{\infty}\frac{3}{4}\delta^{(\frac{3}{2\pi}k)}(\xi) + \frac{1}{8i}\left[\delta^{\left(\frac{3}{2\pi}+\frac{3}{2\pi}k\right)}(\xi) - \delta^{\left(-\frac{3}{2\pi}+\frac{3}{2\pi}k\right)}(\xi)\right] \\ &= \frac{3}{2\pi}\sum_{k=-\infty}^{\infty}\frac{3}{4}\delta^{(\frac{3}{2\pi}k)}(\xi). \qquad (1.14)\end{aligned}$$

The same sampled function is obtained when the function $g(t) \equiv \frac{3}{4}$ is sampled at rate $\lambda = \frac{3}{2\pi}$, so no unique reconstruction of $f^{(ex)}$ is possible.

In particular, high-frequency noise affects the recovery of the low frequency component, in which one is interested. For this reason, often a low-pass filter is applied to the input signal. Such an anti-aliasing filter separates the signal from high-frequency noise and in that way prevents aliasing effects.

Of course, in practice, only sampled values corresponding to a limited time interval can be taken into account, whereas in the above argument, we assumed that the samples are known for all t_n. We will address this issue in the precise mathematical framework of Section 1.2.2. There we will provide an argument that

under certain assumptions, the effect of this practical constraint is of primarily local nature, i.e., after some adjustment interval, it will only lead to a small error.

Formally, one can also compute the frequency domain representation $\widehat{f_\lambda}$ by taking the Fourier transform of the sum in Equation (1.5) term by term. One obtains:

$$\widehat{f_\lambda}(\xi) = \sum_{n=-\infty}^{\infty} f\left(\frac{n}{\lambda}\right) e^{-2\pi i \xi n/\lambda}. \tag{1.15}$$

Although Equation (1.15) is only a formal identity, as the sum will not converge in general, some properties of the Fourier transform (e.g., that convolution in the time domain corresponds to multiplication of the series) will hold true for the formal series. This motivates the definition of the *z-transform*, which is defined to be the formal series in 1.15, where $z = e^{2\pi i \xi/\lambda}$. In mathematics, it is more common to work with the *generating function*, which differs from the z-transform just by a sign in the exponent.

Definition 1.1. For a sequence $(x_n)_{n\in\mathbb{Z}}$, the generating function is the formal series given by

$$X(z) = \sum_{n\in\mathbb{Z}} x_n z^n \tag{1.16}$$

We will usually denote the generating function of a sequence by the corresponding capital letter.

The generating function of the sequence $y_n = \sin(2n)$ of samples of the function $f^{(ex)}$ is $Y(z) = \sum_{n\in\mathbb{Z}} \left(\frac{3}{4} + \frac{1}{4}\sin(2n)\right) z^n$. Again, z is just a formal variable.

1.1.2 Filters in discrete signal processing

One-bit quantization schemes exploit the redundancy that arises in the sequence of sampled values for large sampling rates λ. Consequently, the choice of

each quantized output should depend on more than one sampled value. In order to implement the associated quantization procedure in an analog circuit, one needs to perform operations on the sequence of sampled values that combine samples corresponding to different time instances. To describe these operations, we work with a general ordered sequence $y_n := f\left(\frac{n}{\lambda}\right)$, dropping the explicit reference to the actual sampling time instances. Nevertheless, we will refer to the index n as the "time instance" associated with the sequence element y_n. Using such notation, one has a well-defined notion of the "previous" and the "next" time instance.

A core concept in discrete signal processing is that of a *delay*. A delay element in an analog circuit stores an input; it outputs at each time instance the input it received at the previous time instance. Several delays can be combined to a *linear filter with k tabs*. Applied to an input sequence $(x_n)_{n\in\mathbb{Z}}$, it outputs at time n a linear combination of the inputs x_j corresponding to time instances $n-k$ through $n-1$, where each x_{n-j} is multiplied by a scalar coefficient h_j, which does not depend on the time n. Hence the output u_n is given by

$$u_n = \sum_{j=1}^{k} h_j x_{n-k} = (h * x)_n. \tag{1.17}$$

Most of the time, we require *causality*, i.e., $h_0 = 0$. This is particularly important if the circuit involves a *feedback loop*, i.e., each input x_j depends on the output u_j (or u_l with $l < j$). In this case, the output u of a non-causal filter would be self-referential, thus ill-defined. For example, consider the feedback filter described by the recurrence relation

$$u_n = (h * u)_n. \tag{1.18}$$

If h is the causal sequence given by $h_1 = 1$, $h_2 = -2$ and $h_j = 0$ for all other j

including 0, then Equation (1.18) reads $u_n = u_{n-1} - 2u_{n-2}$, which describes how to compute an output from the previous filter outputs. However, if h is given by $h_0 = 2$, $h_1 = -1$ and thus not causal, then the recurrence relation $u_n = 2u_n - u_{n-1}$ has u_n on both sides of the equation, and one of the inputs of the filter would be its own output.

In both cases, u_n can be found from only a finite number of inputs, and we call h a filter with *finite impulse response* (FIR). In terms of the generating functions, Equation (1.17) corresponds to the multiplicative identity

$$U(z) = H(z)X(z). \tag{1.19}$$

The function $H(z)$ is referred to as the *transfer function* of the filter. The transfer function of a linear filter with k tabs is a polynomial of degree k; if the filter is causal, its constant term is zero. The transfer functions associated with the two example filters h discussed above are given by $H(z) = z - 2z^2$ for the causal and $H(z) = 2 - z^2$ for the non-causal example.

A filter that corresponds to *dividing* the generating function of the input by a polynomial $A = \sum_{j=0}^{k} a_j z^j$ of degree k is also easily implemented using delays. Indeed, if the generating functions of the input y_n and the output u_n satisfy

$$U(z) = \frac{Y(z)}{A(z)}, \tag{1.20}$$

then for the time representations one has

$$y_n = (a * u)_n = \sum_{j=0}^{k} a_j u_{n-j}. \tag{1.21}$$

Denoting $\delta^{(0)}$ the Kronecker delta and normalizing $a_0 = 1$, this equation can be rewritten as

$$u_n = y_n - (a - \delta^{(0)}) * u. \tag{1.22}$$

We have seen that operations of this form can be implemented using a combination of delays. A notable difference between this scenario and the above situation is that here, the input of the filter consists of the previous output, i.e., such a filter is always a feedback filter. Also, the output u_n depends on the input x_n, so the filter is not causal in x. The lack of causality can be resolved by combining this filter with an FIR filter as introduced above, i.e., considering a transfer function of the form $H(z) = \frac{B(z)}{A(z)}$ with $b_0 = 0$ and $a_0 = 1$. One obtains the recurrence relation

$$a * u = b * x \Leftrightarrow u_n = (b * x)_n + ((\delta_0 - a) * u)_n \,, \tag{1.23}$$

and so the combined filter is causal in both x_j and u_j. Such a filter can be implemented using a finite number of delays even in conjunction with a feedback loop.

Note that $H(z)$ is again the generating function of a causal sequence h (it can be obtained by expanding H as a power series), but this sequence will not be finite. We say that such a filter has *infinite impulse response* (IIR).

1.1.3 Quantization

While sampling as described above results in a discretization of the domain, in order to get a true digital representation (i.e., a representation of the signal using finitely many bits), one also needs to discretize the range.

At the core of such an operation is usually a quantizer Q that maps each

component of the input sequence to its sign.

$$Q : x \mapsto q : q_n = sign(x_n) \tag{1.24}$$

Such an ideal quantizer will hardly appear in practice; in particular the sharp cutoff for input values near 0 cannot be realized. Consequently, Daubechies and Devore [4] rigorously showed how their quantization schemes are robust to errors arising from this kind of issues. In this thesis, however, we will assume an ideal quantizer exactly given by Equation (1.24) and leave the discussion of robustness for future work.

While some previous mathematically oriented constructions of one-bit quantization schemes [4] employed a nested sequence of such quantizers, the schemes designed in this thesis will use only one quantizer, as it is common in the engineering literature (for example, [16, 17]) as well as in some mathematically oriented papers on one-bit quantization (for example, [7]).

1.1.4 Noise shaping

In this section, we will show how quantizers of the form (1.24) can be used to design a quantization procedure which efficiently represents a bandlimited signal by a sequence of quantized values q_n from a finite alphabet $\mathcal{A}$. In the same way that the signal can be reconstructed from its samples in Equation (1.13), applying a low-pass filter to the sequence q_n should allow for approximate recovery of the signal. That is, the function

$$\tilde{f}_\lambda(t) = \frac{1}{\lambda} \sum_{n=-\infty}^{\infty} q_n \varphi\left(t - \frac{n}{\lambda}\right) \tag{1.25}$$

should approximate the signal f.

A natural starting point for the design and analysis of such quantization procedures is the sequence $w_n = y_n - q_n$ given by the individual differences between the samples of a bandlimited function and the associated quantized values, often considered to be the "noise" arising from the quantization procedure. If all w_n are small, the signal reconstructed from the quantized values will be close to the original signal. In general, one can only achieve each w_n to be small if one allows for a very large alphabet $\mathcal{A}$. This is the core idea of pulse code modulation as explained in the introduction. In contrast, in this thesis, we will work with the coarse alphabet $\mathcal{A} = \{-1, 1\}$.

If the quantizer (1.24) is applied to each sample independently, however, the result will not lead to a good approximation. For example, the function $f^{(ex)}$ considered above is positive; the resulting sequence q_n of quantized values would just consist of the value 1, which certainly does not sufficiently capture the information. Thus a one-bit quantization procedure must take into account the samples and quantized values corresponding to different time instances as well.

Furthermore, the quantized values should be determined using an on-line procedure: q_n should not explicitly depend on y_j or q_j for $j > n$. Hence the goal must be to base each q_n partly on the previous values q_j for $j < n$ using a feedback loop. Although one can never achieve that all the individual differences w_n are small, one can exploit the fact that a low-pass filter is used for the reconstruction. A *high-pass sequence*, a sequence whose frequency representation is mostly supported for $|\xi|$ large, will be close to the kernel of such a reconstruction operation, so if w is a high-pass sequence, one expects a small reconstruction error. Concretely, this is achieved if $\widehat{w}$ has multiple zeros at $\xi = 0$: this ensures that $\widehat{w}$ is small in a neigh-

borhood of $\xi = 0$, i.e., for the low frequency range. A zero of $\widehat{w}$ at $\xi = 0$, in turn, corresponds to a zero of W at $z = 1$. Accordingly, the quantization procedure should be designed such that the generating function of the noise arising from it takes a particular "shape", with multiple zeros at 1 and most of the mass away from 1. This idea is referred to as *noise shaping*.

A common approach to noise shaping is Sigma-Delta ($\Sigma\Delta$) modulation (also referred to as $\Sigma\Delta$ quantization) as described by Schreier et al. [17, 16, 13], for example. The underlying idea is to apply an appropriate filter to the *quantization noise* defined to be the difference between input and output of the quantizer and feed the result back into the circuit. Figure 1.1 shows the block diagram of a $\Sigma\Delta$ modulator as it is introduced in [17] together with the notation we will be using (which is consistent, for example, with [7]). Here, Q is an ideal quantizer as in Equation (1.24). Due to different conventions, the notation used in [17] differs from ours by a sign. Accordingly, in our notation, the quantization noise is $-v$ and the filter used in the circuit has the coefficient sequence $-h$. In this framework, designing a good $\Sigma\Delta$ modulator amounts to choosing h such that the resulting circuit has good noise shaping properties.

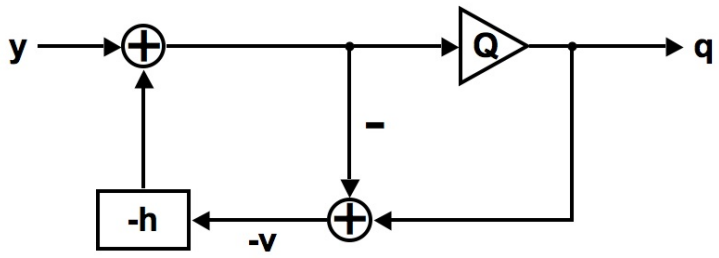

Figure 1.1: Block diagram of a $\Sigma\Delta$ modulator (from [17]), together with the associated notation used in this thesis

The quantization noise arising from the circuit given in Figure 1.1 evolves

according to the recurrence relation

$$v_n = (h * v)_n + y_n - q_n, \tag{1.26}$$

where the quantized values q_n are determined via the non-linear quantization rule

$$q_n = sign((h * v)_n + y_n). \tag{1.27}$$

As this choice of q_n minimizes $|v_n|$ at each time instance, we refer to Equation (1.27) as the *greedy quantization rule*.

In this thesis, we use the term $\Sigma\Delta$ *modulator* also to refer to the dynamical system arising from Recurrence Relation (1.26), together with some rule or procedure (not necessarily (1.27)) that gives rise to the quantized sequence q_n. While we mostly work with the greedy quantization rule in this thesis, we also discuss perturbations of the greedy quantization rule as well as other linear quantization rules. As mentioned above, the $\Sigma\Delta$ modulators designed by Daubechies and DeVore [4] use a completely different quantization rule.

Recurrence Relation (1.26) translates to a condition for w_n:

$$\begin{aligned} w_n &= q_n - y_n \\ &= (h * v)_n - v_n \\ &= (h - \delta_0) * v \end{aligned} \tag{1.28}$$

In terms of the generating functions, this leads to the condition

$$W(z) = (H(z) - 1)V(z) \tag{1.29}$$

Heuristically, we argue that if $H(z) - 1$ has multiple zeros at 1, then so does W; hence w is a high pass sequence for all inputs y. When $H(z) - 1$ has m zeros at 1, we refer to the associated $\Sigma\Delta$ modulator as an *m-th order* modulator.

This heuristic argument can break down if the function $V(z)$ has a pole at 1 or is not even defined in the neighborhood. In particular, the formal factorization (1.29) is possible for all w if one allows v_n to be unbounded. In the circuit, this scenario corresponds to positive feedback: The quantization noise variable will grow over time, and the output will be meaningless. Hence it is crucial that the sequence v_n remains bounded. If this is the case for all possible input signals, we say that the associated $\Sigma\Delta$ modulator is *stable*.

Under the assumption of stability, the heuristic noise shaping arguments can indeed be made rigorous. In Section 1.2.2, we rigorously derive error decay bounds for stable $\Sigma\Delta$ modulators.

1.2 The mathematical perspective

In this section we embed the heuristic arguments of Section 1.1 into a rigorous mathematical framework. We reintroduce several of the concepts mentioned in the previous section, but this time in a precise mathematical setting.

1.2.1 General setup

We define the space of Ω-bandlimited functions to be

$$\mathcal{B}_\Omega = \left\{ \check{\nu}(\xi) = \int_{-\infty}^{\infty} e^{2\pi i x \xi} d\nu \;\middle|\; \nu \text{ is a Borel measure with } supp\,(\nu) \subseteq \left[-\frac{\Omega}{2}, \frac{\Omega}{2}\right] \right\} \tag{1.30}$$

It is easy to see that all of the elements in $\mathcal{B}_\Omega$ are analytic functions. Furthermore, $\mathcal{B}_\Omega$ is contained in the space $\mathcal{S}'$ of tempered distributions (see, for example, [10] for the details of its construction). Since the Fourier transform as an operator on $\mathcal{S}'$ maps $f = \check{\nu} \in \mathcal{B}_\Omega$ to ν, the definition of $\mathcal{B}_\Omega$ is a precise formulation of the intuitive notion from Section 1.1.

Furthermore, recall from Section 1.1 that our model assumption was that the amplitude of the bandlimited function modeling the signal is bounded by some constant μ. For a fixed $\mu \in \mathbb{R}$, the set of all such functions is:

$$\mathcal{B}_\Omega^\mu := \mathcal{B}_\Omega \cap \{f \in L^\infty : \ \|f\|_{L^\infty} \leq \mu\} \tag{1.31}$$

For reasons of notational convenience we will normalize $\Omega = 1$ from now on. The corresponding results for other bandwidths Ω can always be obtained by a suitable rescaling.

A precise version of the Shannon-Nyquist Sampling Theorem given in Equation (1.13) for functions in $\mathcal{B}_1$ was proved in [6]:

Theorem 1.1 ([6])**.** *Let f be in $\mathcal{B}_1$, $\lambda_0 > 1$ and $\varphi \in L^1$ such that*

$$\widehat{\varphi}(\xi) = \begin{cases} 1 \text{ if } |\xi| < 1 \\ \\ 0 \text{ if } |\xi| > \lambda_0 \end{cases} \tag{1.32}$$

Then for all $\lambda \geq \lambda_0$, the following equality holds in the Cesàro mean for all $t \in \mathbb{R}$:

$$f(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} f\left(\frac{n}{\lambda}\right) \varphi\left(t - \frac{n}{\lambda}\right) \tag{1.33}$$

In the sequel, we will assume that $\widehat{\varphi}$ is smooth so that φ is in the Schwartz space

$\mathcal{S}$ and has very strong decay and smoothness properties. Then for $f \in \mathcal{B}_1^\mu$, the series in Equation (1.33) converges absolutely, and one does not need to consider the Cesàro mean in Theorem 1.1.

Property (1.32) captures the nature of a low-pass filter, as described in Section 1.1.1. Thus, Equation (1.33) is a mathematical restatement of the fact that f can be reconstructed exactly by applying a low-pass filter to the sampled function. Here, the filter is modeled by the *low-pass operator*:

$$T_\lambda^\varphi : \ell^\infty(\mathbb{Z}) \to L^\infty(\mathbb{R}) \qquad T_\lambda^\varphi(a) = \frac{1}{\lambda} \sum_{n\in\mathbb{Z}} a_n \varphi\left(t - \frac{n}{\lambda}\right). \tag{1.34}$$

To emphasize the role of the kernel φ, we sometimes use the notation

$$a \circledast_\lambda \varphi := \sum_{n\in\mathbb{Z}} a_n \varphi\left(t - \frac{n}{\lambda}\right) = \lambda T_\lambda^\varphi a. \tag{1.35}$$

This operation can be thought of as a generalized convolution, as suggested by the following lemma:

Lemma 1.2. *For all $a \in \ell^\infty(\mathbb{Z}), b \in \ell^1(\mathbb{Z})$ and $\varphi, \psi \in \mathcal{S}$, the following hold.*

1. $(a * b) \circledast_\lambda \varphi = a \circledast_\lambda (b \circledast_\lambda \varphi)$,
2. $a \circledast_\lambda (\varphi * \psi) = (a \circledast_\lambda \varphi) * \psi$.

$*$ denotes the usual convolution operation for sequences or functions.

Proof. For 1., we calculate

$$(a * b) \circledast_\lambda \varphi(t) = \sum_{n\in\mathbb{Z}} (a * b)_n \varphi\left(t - \frac{n}{\lambda}\right) = \sum_{n\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} a_k b_{n-k} \varphi\left(t - \frac{n}{\lambda}\right)$$

$$= \sum_{k\in\mathbb{Z}} a_k \sum_{n\in\mathbb{Z}} b_{n-k}\varphi\left(t-\frac{n}{\lambda}\right) = \sum_{k\in\mathbb{Z}} a_k \sum_{l\in\mathbb{Z}} b_l \varphi\left(t-\frac{k}{\lambda}-\frac{l}{\lambda}\right)$$
$$= \sum_{k\in\mathbb{Z}} a_k \left(b \circledast_\lambda \varphi\left(t-\frac{k}{\lambda}\right)\right) = a \circledast_\lambda (b \circledast_\lambda \varphi)(t), \tag{1.36}$$

and the second equality follows from

$$a \circledast_\lambda (\varphi * \psi) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} a_n(\varphi*\psi)\left(t-\frac{n}{\lambda}\right) = \sum_{n\in\mathbb{Z}} a_n \int_{\mathbb{R}} \varphi\left(t-\frac{n}{\lambda}-s\right)\psi(s)ds$$
$$= \int_{\mathbb{R}} \sum_{n\in\mathbb{Z}} a_n \varphi\left(t-s-\frac{n}{\lambda}\right)\psi(s)ds = (a \circledast_\lambda \varphi) * \psi. \tag{1.37}$$

Changing the order of summation is justified because the samples of the function φ form an ℓ^1-sequence: φ is in $\mathcal{S}$, hence both Lipschitz and in L^1.

□

1.2.2 Error decay for m-th order $\Sigma\Delta$ modulators

In this section, we discuss the extent to which the signal $\tilde{f}_\lambda$, as reconstructed from the quantized values in Equation (1.25), provides a good approximation to the original signal. The *instantaneous error* at time t is given by the pointwise difference:

$$f(t) - \tilde{f}_\lambda(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} (y_n - q_n)\varphi\left(t-\frac{n}{\lambda}\right). \tag{1.38}$$

There are different approaches to quantifying the quality of the approximation as expressed by this time-dependent error function. In this thesis, we seek to minimize the supremum norm of the pointwise error. Other authors have also considered the L^2-norm [9] or the L^p-norm for general $1 < p < \infty$ [6].

The guiding principle for the analysis in this section shall be the heuristic

considerations of Section 1.1.4. We first recall some concepts from that section:

- A $\Sigma\Delta$ modulator with input sequence y_n is described by the recurrence relation $v_n = (h * v)_n + y_n - q_n$.
- A $\Sigma\Delta$ modulator corresponding to Recurrence Relation (1.26) is an m-th order modulator if the transfer function $H(z)$ associated with the filter h is such that $1-H$ has m zeros at $z = 1$. That is, it should admit a factorization

$$1 - H(z) = (1 - z)^m G(z). \tag{1.39}$$

In order for Relation (1.26) to be well-defined, one needs in addition that $h \in \ell^1$. This constraint is not conveniently expressed in terms of the generating functions; rather one rewrites Equation (1.39) in terms of the associated time representations h and g. One obtains

$$\delta^{(0)} - h = (1, -1) * \cdots * (1, -1) * g, \tag{1.40}$$

where g is such that its generating function is G and $\delta^{(0)}$ denotes the Kronecker delta. Now note that

$$[(1, -1) * u]_n = [\Delta u]_{n-2}, \tag{1.41}$$

where Δ denotes the *finite difference operator* given by $[\Delta u]_n = u_{n+1} - u_n$. To emphasize the convolutional nature of this operator, we sometimes write $\Delta * u$ instead of Δu, with the understanding that Δ is the sequence given by $\Delta_{-1} = 1$, $\Delta_0 = -1$ and $\Delta_j = 0$ for all other j.

The shift of indices in Equation (1.41) often results in different indexing for statements about sequences compared to the corresponding statements about their

generating functions. When we are interested in the ℓ^1-norm, these index shifts do not change the results. For this reason, we will sometimes just state the manipulations that we apply in our derivation up to a shift of indices, without being more specific.

Expressing Equation (1.40) in terms of finite difference operators motivates the following definition:

Definition 1.2. A $\Sigma\Delta$ modulator corresponding to a filter with coefficient sequence h is said to be an m-th order $\Sigma\Delta$ modulator if $\delta_0 - h = \Delta^m g$ for some $g \in \ell^1$.

For the analysis of the dynamical system given by a $\Sigma\Delta$ modulator, Recurrence Relation (1.26) must be considered together with an initial condition. Usually, one works with the initial condition $v_k = 0$ for $k < 0$. This corresponds to the practical constraint that one only has access to the sampled values starting from some initial time instance. Thus, rather than via Equation (1.38), we measure the pointwise error by:

$$e_\lambda(t) = f(t) - \tilde{f}(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{N}} (y_n - q_n) \varphi \left(t - \frac{n}{\lambda} \right) \tag{1.42}$$

and seek to minimize $\|e_\lambda\|_{L^\infty}$. As in [7], we split this quantization error into two parts:

First, we consider the error e_λ^1 arising when we apply the low-pass filter to only the samples corresponding to positive times, $y_n = f\left(\frac{n}{\lambda}\right)$ for $n \geq 0$. This procedure results in the function

$$\bar{f}(t) = \frac{1}{\lambda} \sum_{n=0}^{\infty} y_n \varphi \left(t - \frac{n}{\lambda} \right). \tag{1.43}$$

In order to be able to use the same formalism as above to represent the reconstruction operation, we define, for a sequence $a \in \ell^\infty(\mathbb{N})$, its extension $\bar{a} \in \ell^\infty(\mathbb{Z})$

via

$$\bar{a}_n := \begin{cases} a_n & \text{for } n \geq 0 \\ 0 & \text{for } n < 0. \end{cases} \tag{1.44}$$

In order to rewrite Equation (1.43) in terms of the low-pass operator T_λ^φ, we apply this extension operation to the input sequence y:

$$\bar{f} = T_\lambda^\varphi \bar{y} = \frac{1}{\lambda} \bar{y} \circledast_\lambda \varphi. \tag{1.45}$$

The error arising at time t in this first approximation is

$$e_\lambda^1(t) = \left| \bar{f}(t) - f(t) \right| = \left| \frac{1}{\lambda} \sum_{n \in \mathbb{Z} \setminus \mathbb{N}} f\left(\frac{n}{\lambda}\right) \varphi\left(t - \frac{n}{\lambda}\right) \right|. \tag{1.46}$$

This error will be present independent of the employed quantization procedure. It was shown in [7] that, uniformly in λ, one has

$$\limsup_{t \to \infty} \left| e_\lambda^1(t) \right| = 0. \tag{1.47}$$

Combining this result with a similar estimate for the end point, one can control the error arising from the fact that one has only access to the samples corresponding to a finite-length interval $[0, T_{max}]$, if one disregards an adjustment period of length T large enough at beginning and at the end of the sampling interval. For this reason our analysis will in the following focus on the second component of the error.

The second component of the error arises when $\bar{f}$ is constructed from the sequence of quantized values $q \in \{-1, 1\}^{\mathbb{N}}$. The reconstructed signal $\tilde{f}_\lambda$ can be

expressed as the low-pass operator T_λ^φ applied to the extension $\bar{q}$ as above:

$$\tilde{f}_\lambda = T_\lambda^\varphi \bar{q} = \frac{1}{\lambda} \bar{q} \circledast_\lambda \varphi. \tag{1.48}$$

To find the error that arises in this second approximation step, we need to study how the approximate reconstruction differs from $\bar{f}$. We need to bound the function

$$e_\lambda^2 = \bar{f} - \tilde{f}_\lambda = \frac{1}{\lambda} (\bar{y} - \bar{q}) \circledast_\lambda \varphi. \tag{1.49}$$

Again, we are interested in bounding $\|e_\lambda^2\|_{L^\infty}$ for sequences y that arise as samples $y_n = f\left(\frac{n}{\lambda}\right)$ of a bandlimited signal f as above. To compare the error bounds for different values of λ, one could work with the class of signals $\mathcal{Y}^{(f)} = \left\{y = (y_n)_{n\in\mathbb{N}} : y_n = f\left(\frac{n}{\lambda}\right) \text{ for some } \lambda\right\}$. It is difficult, however, to bound the error in a way that accurately reflects the detailed nature of the signal f. Instead, we note that for $f \in \mathcal{B}_\Omega^\mu$, one has $\mathcal{Y}^{(f)} \subset \mathcal{Y}_\mu$ defined by $\mathcal{Y}_\mu = \{y = (y_n)_{n\in\mathbb{N}} : ||y||_{\ell^\infty} \leq \mu\}$. All of the following considerations work for arbitrary sequences in $\mathcal{Y}_\mu$.

As discussed in Section 1.1.4, stability is an important concept for the error analysis of $\Sigma\Delta$ modulators. Thus it is crucial for the following analysis to give a precise mathematical definition of what we mean by a stable $\Sigma\Delta$ modulator when we work with a specified class of input sequences $\mathcal{Y}$ like for example $\mathcal{Y}_\mu$. The following definition is independent of the quantization rule, we only assume that there is some procedure $\tilde{Q}$ that creates a sequence of quantized values $q \in \{-1, 1\}^\mathbb{N}$ from any input sequence $y \in \mathcal{Y}$.

Definition 1.3. For a fixed causal coefficient sequence $h \in \ell^1$, consider the $\Sigma\Delta$

modulator given by the recurrence relation

$$v_n = (h * v)_n + y_n - q_n, \quad n \in \mathbb{N} \tag{1.50}$$

with the initial condition $v_n = 0$ for $n < 0$ and some arbitrary quantization rule $q = \tilde{Q}(y)$. We say that the modulator is *stable* with respect to a given class $\mathcal{Y}$ of input sequences, if for all y in $\mathcal{Y}$, the sequence v defined recursively by Relation (1.50) is bounded.

The following estimates, from [4] and [7], will provide a bound on $\|e_\lambda^2\|_{L^\infty}$. The proof involves approximation theoretic results from [5]. As the estimates are crucial for the results in this thesis, we include a self-contained presentation of the proof based on these sources.

Theorem 1.3. *For $\varphi \in \mathcal{S}$ and λ_0 fixed as in Theorem 1.1, consider an m-th order $\Sigma\Delta$ modulator and the corresponding sequences g and h as in Definition 1.2. If the modulator is stable for all input sequences $y \in \mathcal{Y}_\mu$, then the decay of the error e_λ^2 can be bounded by*

$$\|e_\lambda^2\|_{L^\infty} \leq \|v\|_{\ell^\infty} \|g\|_{\ell^1} \|\varphi\|_{L^1} \lambda_0^m \pi^m \lambda^{-m}. \tag{1.51}$$

Proof. From Recurrence Relation (1.50) and Lemma 1.2, we obtain:

$$\begin{aligned} e_\lambda &= \frac{1}{\lambda}(\delta^{(0)} - h) * v \circledast_\lambda \varphi = \frac{1}{\lambda} v * (\Delta^m g) \circledast_\lambda \varphi \\ &= \frac{1}{\lambda}(v * g * \Delta^m) \circledast_\lambda \varphi = \frac{1}{\lambda}(v * g) \circledast_\lambda (\Delta^m \circledast_\lambda \varphi) \end{aligned} \tag{1.52}$$

To establish a bound on expression (1.52), we need the following two lemmas.

Lemma 1.4 ([5, 4])**.**

$$(\Delta^m \circledast_\lambda \varphi)\left(t-\frac{n}{\lambda}\right) = \frac{1}{\lambda^{m-1}} \int_{-\infty}^{\infty} \varphi^{(m)}\left(t-\frac{n}{\lambda}+s\right)\left(\chi_{[0,1]}\right)^{*m}(\lambda s)\, ds, \tag{1.53}$$

where f^{*k} denotes the k-fold convolution $f * f * \cdots * f$ (k factors).

Proof. We proceed by induction in m. For $m = 1$, we obtain using the Fundamental Theorem of Calculus

$$\begin{aligned}(\Delta \circledast_\lambda \varphi)\left(t-\frac{n}{\lambda}\right) &= \varphi\left(t-\frac{n}{\lambda}+\frac{1}{\lambda}\right) - \varphi\left(t-\frac{n}{\lambda}\right) \\ = \int_0^{1/\lambda} \varphi'\left(t-\frac{n}{\lambda}+s\right) ds &= \int_{-\infty}^{\infty} \varphi'\left(t-\frac{n}{\lambda}+s\right)\chi_{[0,1]}(\lambda s)\, ds,\end{aligned} \tag{1.54}$$

as desired.

For the induction step, calculate

$$\begin{aligned}&(\Delta^m \circledast_\lambda \varphi)\left(t-\frac{n}{\lambda}\right) \\ &= \Delta \circledast_\lambda \left(\Delta^{m-1} \circledast_\lambda \varphi\right)\left(t-\frac{n}{\lambda}\right) \\ &= \frac{1}{\lambda^{m-2}} \int_{-\infty}^{\infty} \left(\varphi^{(m-1)}\left(t-\frac{n}{\lambda}+s+\frac{1}{\lambda}\right) - \varphi^{(m-1)}\left(t-\frac{n}{\lambda}+s\right)\right)\left(\chi_{[0,1]}\right)^{*(m-1)}(\lambda s)\, ds \\ &= \frac{1}{\lambda^{m-2}} \int_{-\infty}^{\infty} \int_{s}^{s+\frac{1}{\lambda}} \varphi^{(m)}\left(t-\frac{n}{\lambda}+u\right) du \left(\chi_{[0,1]}\right)^{*(m-1)}(\lambda s)\, ds \\ &= \frac{1}{\lambda^{m-2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \chi_{[0,1/\lambda]}(u-s)\varphi^{(m)}\left(t-\frac{n}{\lambda}+u\right)\left(\chi_{[0,1]}\right)^{*(m-1)}(\lambda s)\, ds\, du \\ &= \frac{1}{\lambda^{m-2}} \int_{-\infty}^{\infty} \varphi^{(m)}\left(t-\frac{n}{\lambda}+u\right) \int_{-\infty}^{\infty} \chi_{[0,1]}(\lambda u - v)\left(\chi_{[0,1]}\right)^{*(m-1)}(v)\, \frac{dv}{\lambda} du\end{aligned}$$

$$= \frac{1}{\lambda^{m-1}} \int_{-\infty}^{\infty} \varphi^{(m)} \left(t - \frac{n}{\lambda} + u \right) \left(\chi_{[0,1]} \right)^{*(m)} (\lambda u)\, du \tag{1.55}$$

□

Lemma 1.5 (see for example [5])**.** *For each integer m, the following equality holds a.e. in t:*

$$S(t) := \sum_{j=-\infty}^{\infty} \left(\chi_{[0,1]} \right)^{*m} (t - j) = 1 \tag{1.56}$$

Proof. For $m = 1$, observe that for each $t \notin \mathbb{Z}$, $\chi_{[0,1]}(t - j) = 1$ if and only if $j = \lfloor t \rfloor$ and 0 otherwise. Hence, $\sum_{j=-\infty}^{\infty} \left(\chi_{[0,1]} \right) (t - j) = 1$ a.e.

For $m > 1$, observe that each $\left(\chi_{[0,1]} \right)^{*m} (t - j)$ has compact support, so for each t, the sum S has only finitely many non-zero terms. We will now show that $S'(t) = 0$ in the sense of a distributional derivative. Calculate

$$\begin{aligned} S'(t) &= \sum_{j=-\infty}^{\infty} \chi'_{[0,1]} * \left(\chi_{[0,1]} \right)^{*(m-1)} (t - j) \\ &= \sum_{j=-\infty}^{\infty} \left(\delta^{(0)} - \delta^{(1)} \right) * \left(\chi_{[0,1]} \right)^{*(m-1)} (t - j) \\ &= \sum_{j=-\infty}^{\infty} \left(\left(\chi_{[0,1]} \right)^{*(m-1)} (t - j) - \left(\chi_{[0,1]} \right)^{*(m-1)} (t - j - 1) \right) \qquad (1.57) \\ &= 0 \qquad (1.58) \end{aligned}$$

As above, $\delta^{(x)}$ denotes the Dirac delta function centered at x. In the last step, we use that the expression in (1.57) is a telescoping sum. This shows that S is constant a.e. To find the value of the constant, compute

$$\int_0^1 S(t)dt = \int_0^1 \sum_{j=-\infty}^{\infty} \left(\chi_{[0,1]} \right)^{*m} (t - j)dt$$

$$
\begin{aligned}
&= \sum_{j=-\infty}^{\infty} \int_{0}^{1} \left(\chi_{[0,1]}\right)^{*m} (t-j)dt \\
&= \sum_{j=-\infty}^{\infty} \int_{-j}^{-j+1} \left(\chi_{[0,1]}\right)^{*m} (u)du \\
&= \int_{-\infty}^{\infty} \left(\chi_{[0,1]}\right)^{*m} (u)du \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \chi_{[0,1]}(s) \left(\chi_{[0,1]}\right)^{*(m-1)} (u-s)ds\ du \\
&= \int_{-\infty}^{\infty} \chi_{[0,1]}(s) \int_{-\infty}^{\infty} \left(\chi_{[0,1]}\right)^{*(m-1)} (u-s)du\ ds \\
&= \int_{-\infty}^{\infty} \chi_{[0,1]}(s) \int_{-\infty}^{\infty} \left(\chi_{[0,1]}\right)^{*(m-1)} (v)dv\ ds
\end{aligned} \tag{1.59}
$$

The last expression can be shown to equal 1 using an inductive argument with seed $\int_{-\infty}^{\infty} \chi_{[0,1]} = 1$. This completes the proof, as 1 is the only constant that integrates to 1 on $[0,1]$. □

These two lemmas, together with the fact that $\left(\chi_{[0,1]}\right)^{*m}$ as a convolution of non-negative functions is non-negative, justify the estimate:

$$
\begin{aligned}
|e_\lambda^2(t)| &= \left| \frac{1}{\lambda} (v * g) \circledast_\lambda \left(\Delta^m \circledast_\lambda \varphi(t)\right) \right| \\
&= \frac{1}{\lambda} \left| \sum_{n\in\mathbb{Z}} (v*g)_n \left(\Delta^m \circledast_\lambda \varphi\right) \left(t - \frac{n}{\lambda}\right) \right| \\
&\leq \frac{1}{\lambda} \|v\|_{\ell^\infty} \|g\|_{\ell^1} \left\| \left(\Delta^m \circledast_\lambda \varphi\right) \left(t - \frac{\cdot}{\lambda}\right) \right\|_{\ell^1} \\
&= \frac{1}{\lambda} \|v\|_{\ell^\infty} \|g\|_{\ell^1} \left\| \frac{1}{\lambda^{m-1}} \int_{-\infty}^{\infty} \varphi^{(m)} \left(t - \frac{\cdot}{\lambda} + s\right) \left(\chi_{[0,1]}\right)^{*m} (\lambda s)\, ds \right\|_{\ell^1}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \|v\|_{\ell^\infty} \|g\|_{\ell^1} \frac{1}{\lambda^m} \sum_{n=-\infty}^{\infty} \left| \int_{-\infty}^{\infty} \varphi^{(m)}(u) \left(\chi_{[0,1]}\right)^{*m} (\lambda u - \lambda t + n)\, du \right| \\
&= \|v\|_{\ell^\infty} \|g\|_{\ell^1} \frac{1}{\lambda^m} \int_{-\infty}^{\infty} \left|\varphi^{(m)}(u)\right| \sum_{n=-\infty}^{\infty} \left(\chi_{[0,1]}\right)^{*m} (\lambda u - \lambda t + n)\, du \\
&= \|v\|_{\ell^\infty} \|g\|_{\ell^1} \frac{1}{\lambda^m} \int_{-\infty}^{\infty} \left|\varphi^{(m)}(u)\right| du \\
&= \|v\|_{\ell^\infty} \|g\|_{\ell^1} \left\|\varphi^{(m)}\right\|_{L^1} \lambda^{-m} \qquad (1.60)
\end{aligned}
$$

By Bernstein's inequality, one can bound $\left\|\varphi^{(m)}\right\|_{\ell^1} \leq \lambda_0^m \pi^m \|\varphi\|_{\ell^1}$ (see [12]), which yields

$$
|e_\lambda(t)| \leq \|v\|_{\ell^\infty} \|g\|_{\ell^1} \lambda_0^m \pi^m \|\varphi\|_{L^1} \lambda^{-m}. \qquad (1.61)
$$

Taking the supremum over t, this proves the theorem. □

Remark: In the mathematical literature, the recurrence relation of a ΣΔ modulator is often given in *canonical form*

$$
\Delta^m u_n = y_n - q_n. \qquad (1.62)
$$

Recurrence Relation (1.26) above can be rewritten in canonical form by defining the new variable $u := g * v$. Note that the q_n's are still determined from the original variables v_n using (1.27) or a similar quantizer. In particular, computing q_n from the u_j's may involve the u_j's for all $j < n$. Nevertheless, representing the modulator in canonical form can be useful, as using these variables, the $\|g\|_{\ell^1}$-term in the error bound of Theorem 1.3 is absorbed in the variable u. Indeed, the filter

used in Equation (1.62) is $h_{can} = \Delta^m = \Delta^m \delta^{(0)}$, and one obtains the error bound

$$\|e_\lambda^2\|_{L^\infty} \leq \|u\|_{\ell^\infty} \|\varphi\|_{L^1} \lambda_0^m \pi^m \lambda^{-m}. \tag{1.63}$$

1.2.3 Superpolynomial error decay

The bounds given in Equation (1.61) can only establish polynomial error decay for any given order. Stronger results (cf. [6]) establish better bounds on the error decay for first order schemes, but numerical experiments suggest that for each order, the decay is, nevertheless, polynomial. However, usually both the sampling frequency and the feedback filter of a $\Sigma\Delta$ modulator are built-in parameters of a circuit, and therefore we cannot let $\lambda \to \infty$ in a circuit without changing its architecture. So when we study the asymptotic error decay as $\lambda \to \infty$, what we mean is that, for each λ, we design some modulator M_λ with corresponding error $e_\lambda(M_\lambda)$ and then seek asymptotic bounds for $\|e_\lambda(M_\lambda)\|_{L^\infty}$ as $\lambda \to \infty$. In particular, this procedure can involve choosing schemes of different order for different values of λ. Higher order schemes will have a better asymptotic decay rate but typically involve larger constants. In general, the larger the desired sampling rate λ, the greater the order one should choose.

This approach was first systematically employed by Daubechies and DeVore [4]. They constructed an infinite family of stable $\Sigma\Delta$ modulators, one modulator M_m for each order m. Then, for each sampling frequency they determined an appropriate order $m(\lambda)$ such that $\|e_\lambda\left(M_{m(\lambda)}\right)\|_{\ell^\infty} \to 0$ superpolynomially. They achieved an error decay of order $O\left(\lambda^{-\gamma \log \lambda}\right)$ for some constant γ. The work does not use greedy or linear quantization rules of the type discussed in Sections 1.1.4 and 3.3, but a nested sequence of *sign*-functions.

Güntürk [7] achieved the exponential error decay of $O\left(2^{-0.076\lambda}\right)$ by selecting the scheme of appropriate order from an infinite family, as described in the first paragraph. His work was the first to achieve exponential error decay for one-bit quantization. Section 2.1 provides more details on the underlying construction, based on so-called minimally supported filters. This result has been improved by the author of this thesis to $O\left(2^{-0.087\lambda}\right)$ [11] (This result will not appear in this thesis). Chapter 2 optimizes the error decay in the framework laid out in [7]. The resulting bound for the asymptotic error decay rate is of order $O\left(2^{-0.102\lambda}\right)$. This is the best asymptotic error decay rate currently known for one-bit quantization schemes.

It is known that for any $0 < \mu < 1$, exponential bounds of order $O(2^{-r\lambda})$ are the best possible error decay bounds, which hold uniformly for all input signals $f \in \mathcal{B}_\Omega^\mu$ [3, 7]. More precisely, one has, for any one-bit quantizer:

$$\sup\{\|e_\lambda^2\|_{L^\infty} : f \in \mathcal{B}_\Omega^\mu\} \geq C2^{-\lambda}, \tag{1.64}$$

where C is a constant, which may depend on μ and Ω, but not λ. Thus no exponential error bound with a rate constant $r > 1$ is possible. In [3], even the case $r = 1$ is ruled out.

1.2.4 A basic stability criterion for greedy quantization

For all the previous consideration, stability of the $\Sigma\Delta$ modulator was implicitly assumed. In the engineering literature, stability is often tested for a wide class of input signals (see for example [17]). A rigorous error analysis, however, also requires a rigorous stability analysis. For this, however, not many techniques are

available.

Stability of $\Sigma\Delta$ modulators of order $m = 1$ is immediate (see for example [6]). Yılmaz [19] provided an in-depth stability analysis for certain second order modulators. Daubechies and Devore [4] showed stability for a family of schemes of all orders (compare Section 1.2.3).

The following stability criterion provides a sufficient condition for the stability of a $\Sigma\Delta$-scheme with the greedy quantization rule defined above. It is well-known to the engineering community (see for example [17]), but it was believed to be too restrictive and had not been used for a rigorous stability analysis until Güntürk's work [7].

Theorem 1.6. *Consider a $\Sigma\Delta$ modulator given by the recurrence relation (1.26) with the greedy quantization rule (1.27). If*

$$\|h\|_{\ell^1} \leq 2 - \mu, \tag{1.65}$$

then the modulator is stable for all inputs $y \in \mathcal{Y}_\mu$.

Proof. We prove by induction in n that all $|v_n| \leq 1$. For $n \leq 0$, $|v_n| = 0 \leq 1$ by definition. For $n > 0$, we use the notation $\|w\|_{\ell^\infty}^{(n)} := \sup_{j<n} |w_j|$. Assume that $\|v\|_{\ell^\infty}^{(n)} \leq 1$. Then

$$\begin{aligned}
|v_n| &= |(h * v)_n + y_n - sign\,(h * v + y_n)| \\
&\leq \max\,(1, |(h * v)_n + y_n| - 1) \\
&\leq \max\left(1, \|h\|_{\ell^1}\, \|v\|_{\ell^\infty}^{(n)} + \mu - 1\right) \qquad (1.66) \\
&\leq \max\,(1, (2 - \mu) + \mu - 1) = 1 \qquad (1.67)
\end{aligned}$$

To obtain Line (1.66), we used that h is causal, i.e., $h_n = 0$ for $n \leq 0$. By induction, Inequality (1.67) establishes the theorem. $\square$

Remark: For any $\Sigma\Delta$ modulator of order $m \geq 1$, we have

$$\sum_{j=0}^{\infty}(\delta^{(0)} - h)_j = \sum_{j=0}^{\infty}\Delta\left[\left(\Delta^{m-1}g\right)\right]_j = 0, \tag{1.68}$$

as the second sum is telescoping. Hence $\|h\|_{\ell^1} \geq \|\delta^{(0)}\|_{\ell^1} = 1$.

Recall from Theorem 1.3 that the constant in the error decay bound of a stable m-th order $\Sigma\Delta$ modulator can be bounded in terms of $\|g\|_{\ell^1}$. Thus, to design a modulator that yields the best error bounds, one needs to minimize $\|g\|_{\ell^1}$ over all stable schemes. Since stability can be guaranteed by the criterion given in Theorem 1.6, this motivates, for each m, the following quantitative minimization problem:

$$\text{Minimize } \|g\|_{\ell^1} \text{ subject to } \delta^{(0)} - h = \Delta^m g,\ \|h\|_{\ell^1} \leq 2 - \mu. \tag{1.69}$$

As Theorem 1.6 gives but a sufficient condition for stability, solving this problem is not equivalent to finding the $\Sigma\Delta$ modulators with the best error decay rate. As we will see in Chapter 2, however, even a more restrictive framework will allow for constructions that yield fast error decay.

Chapter 2

Optimizing minimally supported filters

2.1 An optimization problem for filters with minimal support

The results in [7] were based on Minimization Problem (1.69), but they did not provide a complete solution to the problem. Rather, the author introduced a class of feasible filters $h = h^{(m)}$ which were effective in the sense that they led to an exponential error of order $O(2^{-r\lambda})$. These filters $h^{(m)}$ are sparse, i.e., they contain only a few non-zero entries. Indeed, each $h^{(m)}$ has exactly m non-zero entries, which is the minimal support size for which $h^{(m)}$ can satisfy the feasibility conditions: The filter $\delta^{(0)} - h^{(m)}$ arises as the m-th order finite difference of the vector g; therefore its entries have to satisfy m moment conditions. This implies that the support size of $h^{(m)}$ is at least m. We make the following definition:

Definition 2.1. We say that a filter $h = \delta^{(0)} - \Delta^m g$, for a finitely supported g,

has *minimal support* if $|\text{supp } h| = m$.

Here, we require g to be finitely supported, as this is a necessary condition to ensure that both h has finite support and $\|g\|_{\ell^1} < \infty$. The goal of this chapter is to find optimal filters within the class of filters with minimal support.

For filters h with minimal support, i.e.,

$$h = \sum_{j=1}^{m} d_j \delta^{(n_j)}, \tag{2.1}$$

the moment conditions lead to explicit formulae for the entries d_j in terms of the support $\{n_j\}_{j=1}^m$ of h, where $1 \leq n_1 < n_2 < \cdots < n_m$ [7]. Here the condition that $n_1 \geq 1$ follows from the causality of h.

Indeed, one finds

$$d_j = \prod_{i=1}^{m}{}' \frac{n_i}{n_i - n_j}. \tag{2.2}$$

Here the notation $\prod'$, and analogously $\sum'$, indicate that the singular terms are excluded from the product, or the sum respectively. By definition, if $m = 1$, one has $d_1 = 1$.

The condition $\|h\|_{\ell^1} \leq 2 - \mu$ then takes the form

$$\sum_{j=1}^{m} \prod_{i=1}^{m}{}' \frac{n_i}{|n_i - n_j|} \leq 2 - \mu. \tag{2.3}$$

Furthermore, explicit computations lead to the identity

$$\|g\|_{\ell^1} = \frac{\prod_{j=1}^{m} n_j}{m!}. \tag{2.4}$$

In this notation, minimization problem (1.69) takes the form

$$\begin{aligned}&\text{Minimize } \frac{\prod_{j=1}^m n_j}{m!} \text{ over}\\ &\{\mathbf{n}=(n_1,\dots,n_m)\in\mathbb{N}^m : (2.3) \text{ holds and } 1\le n_1<\cdots<n_m\}\end{aligned} \tag{2.5}$$

For $\mu = 1$, problem (2.5) has a solution only for $m = 1$, and we find $h = \delta^{(1)}$, but for $\mu < 1$, the problem has a nontrivial solution for all m. That is, we can find n_j, $j = 1,\dots,m$, that satisfy (2.3). In particular, for $n_j(\sigma) = 1 + \sigma(j-1)$, one shows easily that

$$\lim_{\sigma\to\infty}\sum_{j=1}^m{\prod_{i=1}^m}'\frac{n_i(\sigma)}{|n_i(\sigma)-n_j(\sigma)|}=1. \tag{2.6}$$

So for every $\mu < 1$, $\mathbf{n}(\sigma)$ satisfies constraint (2.3) for all σ large enough.

Furthermore, any minimizer $\mathbf{n}$ of problem (2.5) must satisfy $n_1 = 1$. Indeed, otherwise $n_j > 1$ for all j and we can define $\tilde{\mathbf{n}}$ by $\tilde{n}_j = n_j - 1 \ge 1$ for all $j = 1,\dots,m$. Calculate

$$\sum_{j=1}^m{\prod_{i=1}^m}'\frac{\tilde{n}_i}{|\tilde{n}_i-\tilde{n}_j|}=\sum_{j=1}^m{\prod_{i=1}^m}'\frac{n_i-1}{|n_i-n_j|}<\sum_{j=1}^m{\prod_{i=1}^m}'\frac{n_i}{|n_i-n_j|}\le 2-\mu \tag{2.7}$$

and

$$\frac{\prod_{j=1}^m \tilde{n}_j}{m!}<\frac{\prod_{j=1}^m n_j}{m!}. \tag{2.8}$$

So $\mathbf{n}$ cannot be a minimizer.

Hence, we can fix $n_1 \equiv 1$, which reduces problem (2.5) to minimizing

$$\eta(\mathbf{n}) := \prod_{j=2}^m n_j \tag{2.9}$$

over the set $\{\mathbf{n} = (n_2, \ldots, n_m) \in \mathbb{N}^{m-1} | 1 < n_2 < \cdots < n_m\}$ under the constraint

$$\sum_{j=1}^{m} \prod_{i=1}^{m}{}' \frac{n_i}{|n_i - n_j|} \leq \gamma, \tag{2.10}$$

where again $n_1 \equiv 1$. The factor $m!$ in the denominator has been absorbed into the definition of η to simplify the notation. Furthermore, we have set $\gamma = 2 - \mu$, as the considerations that follow make sense for arbitrary $\gamma > 1$ and not only for $\gamma \leq 2$.

Notational Remark: All quantities in the derivations below depend on m. We will suppress this dependence unless it is relevant in a particular argument.

2.2 The relaxed minimization problem for optimal filters

The variables n_j correspond to the positions of the nonzero entries in the vector h, so they are constrained to positive integer values. We will first consider the *relaxed minimization problem* without this constraint; this will eventually enable us to draw conclusions about the original problem. Thus the variables $n_j \in \mathbb{N}$ will be replaced by relaxed variables $x_j \in \mathbb{R}^+$. Furthermore, it turns out to be convenient to replace the index set $\{1, \ldots, m\}$ by $\{0, \ldots, m-1\}$.

The relaxed minimization problem is specified as follows: Minimize

$$\eta(\mathbf{x}) := \prod_{j=1}^{m-1} x_j \tag{2.11}$$

over the set $D = \{\mathbf{x} \in \mathbb{R}^{m-1} | 1 < x_1 < x_2 < \cdots < x_{m-1}\}$ under the constraint

$$f(\mathbf{x}) := \sum_{j=0}^{m-1} \prod_{i=0}^{m-1}{}' \frac{x_i}{|x_i - x_j|} \leq \gamma, \tag{2.12}$$

where $x_0 \equiv 1$.

Observe that f is defined and smooth in the open domain D. The following monotonicity property for f is important in making inferences from the relaxed to the discrete minimization problem. Let $\mathbf{r}(\mathbf{x})$ be given by $r_j(\mathbf{x}) = \frac{x_j}{x_{j-1}}$, $j = 1, \ldots, m-1$, and set $F(\mathbf{r}) = f(\mathbf{x})$ for $\mathbf{x}$ such that $\mathbf{r} = \mathbf{r}(\mathbf{x})$.

Lemma 2.1. *The function $F(\mathbf{r})$ is strictly decreasing in each variable r_j.*

Proof. A simple calculation shows that

$$F(\mathbf{r}) = \sum_{j=0}^{m-1} \prod_{i=0}^{m-1}{}' \frac{x_i}{|x_i - x_j|} = \sum_{j=0}^{m-1} \prod_{i<j} \frac{1}{r_{i+1} r_{i+2} \cdots r_j - 1} \prod_{i>j} \frac{1}{1 - \frac{1}{r_{j+1} r_{j+2} \cdots r_i}}, \tag{2.13}$$

from which the monotonicity is immediate. □

Definition 2.2. If $\mathbf{x}, \mathbf{y} \in D$ and $1 \leq \frac{y_1}{x_1} \leq \cdots \leq \frac{y_m}{x_m}$, we say that $\mathbf{y}$ is *subordinate* to $\mathbf{x}$.

Clearly, $\mathbf{y}$ is subordinate to $\mathbf{x}$ if and only if $r_j(\mathbf{x}) \leq r_j(\mathbf{y})$ for $j = 1, \ldots, m-1$, so Lemma 2.1 is equivalent to the following:

Corollary 2.2. *If $\mathbf{y}$ is subordinate to $\mathbf{x}$ and $\mathbf{x} \neq \mathbf{y}$, then $f(\mathbf{y}) < f(\mathbf{x})$.*

If $\mathbf{x}$ is a minimizer of the constraint optimization problem (2.11), (2.12), then $f(\mathbf{x}) = \gamma$. Indeed, for a proof by contradiction, assume that $\mathbf{x}$ is a minimizer and $f(\mathbf{x}) < \gamma$. Then for $t \in [0, 1)$, we can define $\tilde{x}_j(t) = (1-t)x_j + tx_0$. Since $f \circ \tilde{\mathbf{x}}$ is continuous in t, and

$$f(\tilde{\mathbf{x}}(0)) = f(\mathbf{x}) < \gamma. \tag{2.14}$$

there exists $t > 0$ such that $f(\tilde{\mathbf{x}}(t)) < \gamma$. However, the function

$$\eta(\tilde{\mathbf{x}}(t)) = \prod_{j=0}^{m-1} ((1-t)x_j + tx_0)) \tag{2.15}$$

is decreasing in t, so

$$\eta(\tilde{\mathbf{x}}(t)) < \eta(\mathbf{x}), \tag{2.16}$$

and $\mathbf{x}$ cannot be a minimizer. Hence we can replace constraint (2.12) by the equality

$$f(\mathbf{x}) = \sum_{j=0}^{m-1} \prod_{i=0}^{m-1}{}' \frac{x_i}{|x_i - x_j|} = \gamma. \tag{2.17}$$

As we now show, this equation defines a smooth manifold within D. It is enough to verify that $\nabla f \neq 0$. Note first that

$$\frac{\partial}{\partial x_k} \frac{x_k}{|x_k - x_j|} = -x_j \frac{1}{x_k - x_j} \frac{1}{|x_k - x_j|}. \tag{2.18}$$

Now calculate for $j \neq k$ using this fact

$$\begin{aligned} \frac{\partial}{\partial x_k} \prod_{i=0}^{m-1}{}' \frac{x_i}{|x_i - x_j|} &= \left(\prod_{\substack{i=0 \\ i \neq k}}^{m-1}{}' \frac{x_i}{|x_i - x_j|} \right) \left(-x_j \frac{1}{x_k - x_j} \frac{1}{|x_k - x_j|} \right) \\ &= -\frac{\eta(\mathbf{x})}{x_k} \frac{(-1)^j}{x_k - x_j} b_j, \end{aligned} \tag{2.19}$$

where we set from now on

$$b_j(\mathbf{x}) = \prod_{i=0}^{m-1}{}' \frac{1}{x_i - x_j}. \tag{2.20}$$

Note that $(-1)^j b_j(\mathbf{x})$ is always positive.

Furthermore, for $j = k$,

$$\frac{\partial}{\partial x_k} \prod_{i=0}^{m-1}{}' \frac{x_i}{|x_i - x_k|} = -\sum_{l=0}^{m-1}{}' \frac{\eta(\mathbf{x})}{x_k} \frac{(-1)^k}{x_k - x_l} b_k(\mathbf{x}). \tag{2.21}$$

Hence

$$\frac{\partial f}{\partial x_k} = -\frac{1}{x_k}\eta(\mathbf{x}) \sum_{j=0}^{m-1}{}' \frac{1}{x_k - x_j} \left((-1)^k b_k(\mathbf{x}) + (-1)^j b_j(\mathbf{x})\right). \tag{2.22}$$

For $k = m - 1$, all terms in the sum are positive. Hence

$$\frac{\partial f}{\partial x_{m-1}} < 0 \tag{2.23}$$

and so $\{\mathbf{x} : f(\mathbf{x}) = \gamma\}$ is a manifold within D.

We now show that the infimum of η subject to (2.17) is attained in D. Let $\eta_0 = \inf_{\mathbf{x}\in D, f(\mathbf{x})=\gamma} \eta(\mathbf{x})$ and let $\mathbf{x}^{(n)} \in D \cap \{f = \gamma\}$ be chosen such that $\lim_{n\to\infty} \eta(\mathbf{x}^{(n)}) = \eta_0$. As before, we set $x_0^{(n)} \equiv 1$.

We first show that $\mathbf{x}^{(n)}$ is bounded. Define $M := \sup_{n\in\mathbb{N}} \eta\left(\mathbf{x}^{(n)}\right)$. Then for each n,

$$\|\mathbf{x}^{(n)}\|_{\ell^\infty} = |\mathbf{x}_{m-1}^{(n)}| \le \eta\left(\mathbf{x}^{(n)}\right) \le M, \tag{2.24}$$

as, for each i, $1 \le \mathbf{x}_i^{(n)} \le \mathbf{x}_{m-1}^{(n)}$. Since $M < \infty$, it follows that $\mathbf{x}^{(n)}$ is bounded. We conclude that $\mathbf{x}^{(n)}$ must have a convergent subsequence $\mathbf{x}^{(n_k)} \to \mathbf{x}^{(\infty)}$.

Now $\mathbf{x}^{(\infty)}$ cannot lie on the boundary of D. Indeed, for any $0 \le j \ne k \le m-1$, we have

$$\gamma = f(\mathbf{x}^{(n)}) \ge \prod_{i=0}^{m-1}{}' \frac{x_i^{(n)}}{|x_i^{(n)} - x_j^{(n)}|} \ge \frac{1}{M^{m-2}|x_j^{(n)} - x_k^{(n)}|}, \tag{2.25}$$

which implies that $|x_j^{(n)} - x_k^{(n)}| \geq \frac{1}{\gamma M^{m-2}} > 0$. It follows that $\mathbf{x}^{(n)}$ stays away from the boundary of D, which implies that $\mathbf{x}^{(\infty)} \in D$. Thus, problem (2.11), (2.17) must have at least one minimizer $\mathbf{x}_{min} = \mathbf{x}^{(\infty)}$ in D. Note that a priori, there can be more than one minimizer.

As $\{\mathbf{x}|f(\mathbf{x}) = \gamma\}$ is a manifold within D, every minimizer $\mathbf{x}_{min} = (x_1, \ldots, x_{m-1})$ of the constrained optimization problem given by (2.11) and (2.17) solves the associated Lagrange multiplier equations, i.e., there exists $\nu = \nu(\mathbf{x}_{min}) \in \mathbb{R}$ such that

$$\nu \nabla \eta(\mathbf{x}_{min}) + \nabla f(\mathbf{x}_{min}) = 0, \tag{2.26}$$

$$f(\mathbf{x}_{min}) = \gamma. \tag{2.27}$$

Combined with (2.22) and the relation $\frac{\partial}{\partial y_k}\eta(\mathbf{y}) = \frac{1}{y_k}\eta(\mathbf{y})$, the Lagrange multiplier equations (2.26), (2.27) take the explicit form

$$\sum_{j=0}^{m-1}{}' \frac{1}{x_k - x_j}\left((-1)^k b_k(\mathbf{x}_{min}) + (-1)^j b_j(\mathbf{x}_{min})\right) = \nu, \tag{2.28}$$

$$f(\mathbf{x}_{min}) = \gamma \tag{2.29}$$

for $k = 1, \ldots, m-1$ and $x_0 \equiv 0$ as before.

Note that any critical point $\mathbf{x}_{crit}$ of the minimization problem for η on D solves equations (2.28), (2.29) for some ν. In the Section 2.4, we will show that in fact η has a unique critical point in D. Before that, we recall some results about Chebyshev Polynomials, which are used in the proof.

2.3 Some useful properties of Chebyshev polynomials

Recall that the Chebyshev Polynomials of the first and second kind in $x = \cos\theta$ are given by

$$T_m(x) = \cos m\theta \qquad \text{and} \qquad U_m(x) = \frac{\sin(m+1)\theta}{\sin\theta}, \tag{2.30}$$

respectively. The Chebyshev polynomials have, in particular, the following properties (see [18], [2]):

- $T_m'(x) = mU_{m-1}(x)$,
- The zeros of U_{m-1} are $z_j = \cos\left(\frac{m-j}{m}\pi\right)$, $j = 1, \ldots, m-1$,
- For $m > 0$, the leading coefficient of T_m is 2^{m-1},
- The Chebyshev polynomials satisfy the following identities

$$T_m(\cosh\tau) = \cosh(m\tau), \quad U_m(\cosh\tau) = \frac{\sinh(m\tau)}{\sinh\tau}, \tag{2.31}$$

- The Chebyshev polynomials satisfy the differential equation

$$(1-x)^2 T_m''(x) - xT_m'(x) + m^2 T_m(x) = 0. \tag{2.32}$$

We say that a polynomial p of degree m has the *equi-oscillation property* on $[-1, 1]$ (compare [2]) if it has $m-1$ real critical points $\zeta_1, \ldots, \zeta_{m-1}$ which satisfy

$$\zeta_0 := -1 < \zeta_1 < \cdots < \zeta_{m-1} < \zeta_m := 1 \tag{2.33}$$

such that the associated values are alternating

$$p(\zeta_j) = (-1)^{m-j} \tag{2.34}$$

for $j = 0, \ldots, m$.

Note that if a polynomial has the equi-oscillation property then its leading coefficient is positive. The Chebyshev polynomials of the first kind T_m have the equi-oscillation property for all m. Indeed, the first two properties given above imply that the z_j's are the critical points of T_m, and a simple calculation shows that $T_m(z_j) = (-1)^{m-j}$. The equi-oscillation property in fact characterizes the Chebyshev polynomials of the first kind:

Proposition 2.3. *If $p(s)$ is a polynomial of degree m in s with the equi-oscillation property on $[-1, 1]$, then $p = T_m$.*

Proof. The proof follows ideas used in [2] to establish that, up to a constant, the T_m are the unique monic polynomials with minimal L^∞ norm.

Let $p(s) = a_p s^m + \ldots$ and $q(s) = a_q s^m + \ldots$ be two polynomials with the equi-oscillation property. W.l.o.g. assume $a_q \geq a_p > 0$. Let $\zeta_1 < \cdots < \zeta_{m-1}$ be the critical points of p in $[-1, 1]$ and set $\zeta_0 = -1$, $\zeta_m = 1$.

Consider the polynomial $r(s) = p(s) - \frac{a_p}{a_q} q(s)$ of degree $(m-1)$. Then $r(\zeta_{m-j}) \geq 0$ for all even j, and $r(\zeta_{m-j}) \leq 0$ for all odd j. The proof that $r \equiv 0$ follows from the following more general statement:

Claim: *If $t_0 < t_1 < \cdots < t_m \in \mathbb{R}$ and a polynomial ρ of degree $m-1$ satisfies $(-1)^j \rho(\zeta_j) \geq 0$ for all j, then $\rho \equiv 0$.*

Proof. The proof proceeds by induction in m. In the case $m = 1$, $\rho(t_0) \geq 0$ and $\rho(t_1) \leq 0$ implies that $r \equiv 0$. For the induction step, assume that the claim holds

true for m. Given a polynomial ρ of degree m with the property, it must have a zero z with $t_m \leq z \leq t_{m+1}$. Define $\tilde{\rho}(x) = \frac{r(x)}{z-x}$. Note that $\tilde{\rho}$ is a polynomial of degree $m-1$. If $z > t_m$, then $\tilde{\rho}(t_j)(-1)^j \geq 0$ for $0 \leq j \leq m$ and hence $\tilde{\rho} \equiv 0$ by the induction hypothesis. If $z = t_m$ then $\tilde{\rho}(t_j)(-1)^j \geq 0$ for $0 \leq j \leq m-1$, but clearly one also has $\tilde{\rho}(t_{m+1})(-1)^m \geq 0$. Again by the induction hypothesis $\tilde{\rho} \equiv 0$. □

We conclude that $r = p - \frac{a_p}{a_q} q \equiv 0$ by applying the claim to $\rho = (-1)^m r$. Since $p(1) = q(1) = 1$ implies that $a_p = a_q$, we see that $p \equiv q$. □

The following lemma plays a useful role in solving the relaxed minimization problem.

Lemma 2.4. *Let z_j, $j = 1, \ldots, m-1$, be the critical points of the Chebyshev polynomial of the first kind T_m, as above, and set $z_0 \equiv -1$. Then*

$$\prod_{\substack{i=0 \\ i \neq k}}^{m-1} (z_k - z_i) = \begin{cases} \frac{m(-1)^{m-1}}{2^{m-1}} & \text{for } k = 0 \\ \\ \frac{m(-1)^{m-1-k}}{2^{m-1}(1-z_k)} & \text{for } k > 0 \end{cases} \tag{2.35}$$

Proof. Recall that T_m has leading coefficient 2^{m-1}. We obtain

$$T'_m(z) = m 2^{m-1} \prod_{i=1}^{m-1} (z - z_i), \tag{2.36}$$

and

$$T''_m(z) = m 2^{m-1} \sum_{j=1}^{m-1} \prod_{\substack{i=1 \\ i \neq j}}^{m-1} (z - z_i), \tag{2.37}$$

and hence for $1 \leq k \leq m-1$

$$T_m''(z_k) = m2^{m-1} \prod_{\substack{i=1 \\ i \neq k}}^{m-1} (z_k - z_i)$$
$$= \frac{m2^{m-1}}{1+z_k} \prod_{\substack{i=0 \\ i \neq k}}^{m-1} (z_k - z_i). \tag{2.38}$$

Thus for $1 \leq k \leq m-1$ one has $T_m'(z_k) = 0$ and $T_m(z_k) = (-1)^{m-k}$, and so (2.32) reads

$$(1-z_k^2)\frac{m2^{m-1}}{1+z_k} \prod_{\substack{i=0 \\ i \neq k}}^{m-1} (z_k - z_i) + m^2(-1)^{m-k} = 0, \tag{2.39}$$

or

$$\prod_{\substack{i=0 \\ i \neq k}}^{m-1} (z_k - z_i) = \frac{(-1)^{m-k-1}m}{2^{m-1}(1-z_k)}. \tag{2.40}$$

On the other hand, as $z_0 = \cos(\pi)$, we have using (2.30)

$$\prod_{i=1}^{m-1} (z_0 - z_i) = \frac{1}{m2^{m-1}} T_m'(z_0) = \frac{1}{2^{m-1}} U_m(z_0) = \frac{1}{2^{m-1}} \lim_{\theta \to \pi} \frac{\sin(m\theta)}{\sin\theta} = \frac{(-1)^{m-1}m}{2^{m-1}}. \tag{2.41}$$

□

2.4 Solution of the relaxed minimization problem

Theorem 2.5. *The minimum value of η on the manifold $\{f = \gamma\}$ in D is given by*

$$\eta = \eta_{min} = \frac{\sinh(2m\beta)}{(2\sinh\beta)^{2m-1}\cosh\beta} \tag{2.42}$$

where $\beta = \beta(m, \gamma)$ is the unique positive solution of the equation

$$\frac{\cosh\left((2m-1)\beta\right)}{\cosh\beta} = \gamma. \tag{2.43}$$

The minimum value η_{min} is attained at the unique point $\mathbf{x}_{min} = (x_1, \ldots, x_{m-1})$, where

$$x_j = 1 + \frac{1}{2\sinh^2\beta}\left(1 + z_j\right), \quad j = 1, \ldots, m-1. \tag{2.44}$$

Here $z_j = cos\left(\frac{m-j}{m}\pi\right)$, $j = 1, \ldots, m-1$, are the zeros of the Chebyshev polynomial of the second kind of degree $m-1$.

Proof. The minimization problem (2.11), (2.17) assumes its minimum in D, so there must be at least one critical point $\mathbf{x}_{crit} = (x_1, \ldots, x_{m-1})$ with $1 < x_1 < \cdots < x_{m-1}$.

To prove uniqueness, we will express the associated Lagrange multiplier problem as a nonlinear matrix equation and then show using a rank argument, which is established by Proposition 2.6, that the equation can have only the solution given by (2.44).

As in (2.28), $\mathbf{x}_{crit} = (x_1, \ldots, x_{m-1})$ must satisfy

$$\sum_{j=0}^{m-1}{}' \frac{1}{x_k - x_j}\left((-1)^k b_k(\mathbf{x}_{crit}) + (-1)^j b_j(\mathbf{x}_{crit})\right) = \nu(\mathbf{x}_{crit}), \tag{2.45}$$

for $k = 1, \ldots, m$ and, again, $b_j(\mathbf{x}_{crit}) = \prod_{i=0}^{m-1}{}' \frac{1}{x_i - x_j}$.

In matrix notation, the statement reads

$$B(\mathbf{x}_{crit})\mathbf{v} = \nu(\mathbf{x}_{crit})\mathbf{e}, \tag{2.46}$$

where $\mathbf{e} = (1, 1, \ldots, 1)^T \in \mathbb{R}^{m-1}$, $\mathbf{v} = (1, -1, 1, -1, \ldots)^T \in \mathbb{R}^m$ and the matrix-valued function $B : \mathbb{R}^{m-1} \to \mathbb{R}^{(m-1)\times m}$ is given by

$$B(\mathbf{y}) = \begin{pmatrix} \frac{b_0(\mathbf{y})}{y_1-y_0} & \sum_{j=0}^{m-1}{}' \frac{b_1(\mathbf{y})}{y_1-y_j} & \frac{b_2(\mathbf{y})}{y_1-y_2} & \cdots & \frac{b_{m-1}(\mathbf{y})}{y_1-y_{m-1}} \\ \frac{b_0(\mathbf{y})}{y_2-y_0} & \frac{b_1(\mathbf{y})}{y_2-y_1} & \sum_{j=0}^{m-1}{}' \frac{b_2(\mathbf{y})}{y_2-y_j} & \cdots & \frac{b_{m-1}(\mathbf{y})}{y_2-y_{m-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{b_0(\mathbf{y})}{y_{m-1}-y_0} & \frac{b_1(\mathbf{y})}{y_{m-1}-y_1} & \frac{b_2(\mathbf{y})}{y_{m-1}-y_2} & \cdots & \sum_{j=0}^{m-1}{}' \frac{b_{m-1}(\mathbf{y})}{y_{m-1}-y_j} \end{pmatrix}, \tag{2.47}$$

where $\mathbf{y} = (y_1, \ldots, y_{m-1})$ and as before $y_0 \equiv 1$.

For given $\mathbf{y} = (y_1, \ldots, y_{m-1})$ let $p_{\mathbf{y}}(s)$ be a polynomial such that

$$p'_{\mathbf{y}}(s) = \prod_{j=1}^{m-1}(s - y_j). \tag{2.48}$$

For definiteness, we normalize $p_{\mathbf{y}}(0) = 0$. Let Γ be a positively oriented circle in $\mathbb{C}$ of radius R large enough to enclose all y_j's, including $y_0 \equiv 1$. We now calculate the integral

$$I_k = \frac{1}{2\pi i}\oint_{\Gamma} \frac{p_{\mathbf{y}}(z)}{(z-y_k)(z-y_0)p'_{\mathbf{y}}(z)}dz, \quad k = 1, \ldots, m-1 \tag{2.49}$$

in two different ways.

Firstly, letting $R \to \infty$, we see that $I_k = \frac{1}{m}$. Secondly, we compute the integral using the residues at y_j, $0 \le j \le m-1$. For the residue R_j at y_j, $j \neq k$, we obtain

$$R_j = (-1)^{m-1}\frac{b_j(\mathbf{y})}{y_j - y_k}p(y_j). \tag{2.50}$$

At z_k, we have a double root in the denominator of the integrand in (2.49), so

$$R_k = \left(\frac{p_{\mathbf{y}}(z)}{\prod\limits_{\substack{i=0 \\ i\neq k}}^{m-1} (z - y_i)} \right)' \Bigg|_{z=y_k} = \frac{p'_{\mathbf{y}}(y_k)}{\prod\limits_{\substack{i=0 \\ i\neq k}}^{m-1} (y_k - y_i)} - p_{\mathbf{y}}(y_k) \sum_{\substack{j=0 \\ j\neq k}}^{m-1} \frac{\prod\limits_{\substack{i=0 \\ i\neq j,k}}^{m-1} (y_k - y_i)}{\left(\prod\limits_{\substack{i=0 \\ i\neq k}}^{m-1} (y_k - y_i) \right)^2}$$

$$= (-1)^{m-1} \sum_{j=0}^{m-1}{}' \frac{b_k(\mathbf{y})}{y_j - y_k} p_{\mathbf{y}}(y_k) \tag{2.51}$$

Summing the residues, we conclude that for $k = 1, \ldots, m-1$:

$$\frac{(-1)^{m-1}}{m} = \sum_{j=0}^{m-1}{}' \frac{1}{y_j - y_k} \left(b_k(\mathbf{y}) p_{\mathbf{y}}(y_k) + b_j(\mathbf{y}) p_{\mathbf{y}}(y_j) \right) \tag{2.52}$$

or equivalently

$$B(\mathbf{y}) \mathbf{p}_{\mathbf{y}} = \frac{(-1)^m}{m} \mathbf{e}, \tag{2.53}$$

where $\mathbf{p}_{\mathbf{y}} = (p_{\mathbf{y}}(y_0), p_{\mathbf{y}}(y_1), \ldots, p_{\mathbf{y}}(y_{m-1})$.

The normalization $p_{\mathbf{y}}(0) = 0$ plays no role in the above calculation, and so (2.53) also holds for $\mathbf{p}_{\mathbf{y}} + \mathbf{e}$. Hence, the vector $\mathbf{e}$ lies in the kernel of $B(\mathbf{y})$ for any $\mathbf{y}$. In Proposition 2.6, we will show that $\dim \operatorname{Ker} B(\mathbf{y}) = 1$, and hence $\operatorname{Ker} B(\mathbf{y})$ is spanned by $\mathbf{e}$. In particular, this shows that $\nu(\mathbf{x}_{crit}) \neq 0$: Otherwise, v would be collinear to e, which is impossible.

Specifying $\mathbf{y} = \mathbf{x}_{crit}$, one obtains

$$B(\mathbf{x}_{crit}) \mathbf{p}_{\mathbf{x}_{crit}} = \frac{(-1)^m}{m} \mathbf{e}, \tag{2.54}$$

and it follows that

$$B(\mathbf{x}_{crit})\left[m(-1)^m\nu(\mathbf{x}_{crit})\mathbf{p}_{\mathbf{x}_{crit}}\right] = \nu(\mathbf{x}_{crit})\mathbf{e}. \tag{2.55}$$

By (2.46), $\mathbf{v}$ also solves (2.55), and thus

$$\mathbf{v} - m(-1)^m\nu(\mathbf{x}_{crit})\mathbf{p}_{\mathbf{x}_{crit}} = c\,\mathbf{e} \tag{2.56}$$

for some constant c.

Set

$$q(s) = m(-1)^m\nu(\mathbf{x}_{crit})p_{\mathbf{x}_{crit}}(s) + c. \tag{2.57}$$

Then q is a polynomial of degree m with critical points at the x_j, $j = 1, \ldots, m-1$ such that $q(x_j) = (-1)^j$, $j = 1, \ldots, m-1$. As q cannot have any more critical points and must be monotonic for $x > x_{m-1}$, then ultimately, it will change sign and there is a unique point $x_m > x_{m-1}$ such that $q(x_m) = -q(x_{m-1}) = (-1)^m$. Hence the polynomial u given by

$$u(s) = (-1)^m q\left(\frac{x_m - 1}{2}(s+1) + 1\right) \tag{2.58}$$

has the equi-oscillation property, and we conclude by Proposition 2.3 that $u(s) = T_m(s)$. That implies, that if z_m, $j = 1, \ldots, m-1$, are the extrema of T_m – i.e., the zeros of the Chebyshev polynomials of the second kind of degree $m-1$ – then the extrema of q are given by

$$x_j = \frac{x_m - 1}{2}(1 + z_j) + 1. \tag{2.59}$$

Thus all critical points $\mathbf{x}_{crit} = (x_1, \ldots, x_{m-1})$ are given by

$$x_j = x_j(K) := 1 + K(1 + z_j) \tag{2.60}$$

for some constant K. It follows from Lemma 2.1 that $f(\mathbf{x}(K))$ is strictly monotonic in K, i.e., different values of K correspond to different values of γ. This proves that η has a unique critical point on $\{f = \gamma\}$ in D, and, in particular, that $\mathbf{x}_{min}$ is unique and given by (2.44).

We now compute $K = K(m, \gamma)$. The calculation uses several facts about Chebyshev polynomials and their roots from Section 2.3. Let $\beta = \beta(m, \gamma) > 0$ be defined through the relation

$$K = \frac{1}{2\sinh^2 \beta}. \tag{2.61}$$

Noting that $z_i = -z_{m-i}$, we obtain from Lemma 2.4

$$\begin{aligned}
f(\mathbf{x}) = & \prod_{i=1}^{m-1} \frac{1 + K(1 + z_i)}{K\,|z_i - z_0|} + \sum_{j=1}^{m-1} \prod_{\substack{i=0 \\ i \neq j}}^{m-1} \frac{1 + K(1 + z_i)}{K\,|z_i - z_j|} \\
= & \frac{2^{m-1}}{m} \prod_{i=1}^{m-1} \frac{1 + K(1 + z_i)}{K} + \sum_{j=1}^{m-1} \frac{2^{m-1}(1 - z_j)}{m} \prod_{\substack{i=0 \\ i \neq j}}^{m-1} \frac{1 + K(1 + z_i)}{K} \\
= & \left[\frac{2^{m-1}}{m} \prod_{i=1}^{m-1} \left(\frac{1}{K} + 1 - z_{m-i} \right) \right] \left[1 + \sum_{j=1}^{m-1} \frac{1 + z_{m-j}}{1 + K(1 - z_{m-j})} \right] \\
= & \left[\frac{2^{m-1}}{m} \prod_{i=1}^{m-1} \left(\frac{1}{K} + 1 - z_i \right) \right] \left[1 + \frac{1}{K} \sum_{j=1}^{m-1} \frac{1 + z_j}{\left(1 + \frac{1}{K}\right) - z_j} \right].
\end{aligned} \tag{2.62}$$

Now $1+\frac{1}{K} = 1+2\sinh^2(\beta) = \cosh(2\beta)$ and

$$\frac{2^{m-1}}{m}\prod_{i=1}^{m-1}\left(\frac{1}{K}+1-z_i\right) = \frac{1}{m^2}T'_m\left(1+\frac{1}{K}\right) = \frac{1}{m^2}T'_m\left(\cosh(2\beta)\right) = \frac{\sinh(2m\beta)}{m\sinh(2\beta)}. \tag{2.63}$$

Furthermore, differentiating (2.31), we obtain for $z=\cosh(\tau)$

$$\sum_{j=1}^{m-1}\frac{1}{z-z_j} = \frac{T''_m(z)}{T'_m(z)} = \frac{m\coth(m\tau)-\coth(\tau)}{\sinh\tau} \tag{2.64}$$

Hence

$$\begin{aligned}
&1+\frac{1}{K}\sum_{j=1}^{m-1}\frac{1+z_j}{\left(1+\frac{1}{K}\right)-z_j}\\
&= 1+(\cosh(2\beta)-1)\sum_{j=1}^{m-1}-1+\frac{1+\cosh(2\beta)}{\cosh(2\beta)-z_j}\\
&= 1+(m-1)(\cosh(2\beta)-1)+(\cosh^2(2\beta)-1)\sum_{j=1}^{m-1}\frac{1}{\cosh(2\beta)-z_j}\\
&= 1-(m-1)(\cosh(2\beta)-1)+\sinh^2(2\beta)\frac{m\coth(2m\beta)-\coth(2\beta)}{\sinh 2\beta}\\
&= m\left(1-\cosh(2\beta)+\sinh(2\beta)\coth(2m\beta)\right)
\end{aligned} \tag{2.65}$$

Combining (2.65) and (2.63) yields

$$\begin{aligned}
\gamma = f(\mathbf{x}) &= \frac{\sinh(2m\beta)-\cosh(2\beta)\sinh(2m\beta)+\sinh(2\beta)\cosh(2m\beta)}{\sinh(2\beta)}\\
&= \frac{\sinh(2m\beta)-\sinh((2m-2)\beta)}{2\sinh(\beta)\cosh(\beta)}\\
&= \frac{2\cosh((2m-1)\beta)\sinh(\beta)}{2\sinh(\beta)\cosh(\beta)}\\
&= \frac{\cosh((2m-1)\beta)}{\cosh(\beta)},
\end{aligned} \tag{2.66}$$

which proves (2.43). As $\frac{\cosh((2m-1)\beta)}{\cosh(\beta)}$ is strictly monotonic in β, $\beta > 0$ is uniquely determined from γ. Of course, this fact also follows from the uniqueness of K proved above.

Now finally, using (2.63), we write

$$\eta_{min} = \prod_{i=0}^{m-1}(1 + K(1 + z_i)) = K^{m-1}\prod_{i=0}^{m-1}\left(\frac{1}{K} + 1 + z_i\right) = \frac{\sinh(2m\beta)}{(2\sinh(\beta))^{2m-1}\cosh(\beta)} \tag{2.67}$$

□

It remains to show that $B(\mathbf{x}_{crit})$ has rank $m-1$. We will show, more generally, that $B(\mathbf{y})$ has rank $m-1$ for an arbitrary $\mathbf{y} = (y_1, \ldots, y_{m-1})$, as long as $y_i \neq y_j$ for $i \neq j$ in $\{0, 1, \ldots, m-1\}$. As before we set $y_0 \equiv 1$. The proof of Proposition 2.6 below goes through without this restriction on y_0, but this more general fact is of no consequence for the results in this paper.

Factor out $b_j(\mathbf{y})$ from the j-th column, $j = 0, \ldots m-1$ and extend the resulting matrix to an $m \times m$ square matrix $\tilde{B}(\mathbf{y})$ by adding a row that is the negative of the sum of all the other rows, as follows.

$$\tilde{B}(\mathbf{y}) = \begin{pmatrix} \sum_{l=0}^{m-1}{}' \frac{1}{y_0-y_l} & \frac{1}{y_0-y_1} & \frac{1}{y_0-y_2} & \cdots & \frac{1}{y_0-y_{m-1}} \\ \frac{1}{y_1-y_0} & \sum_{l=0}^{m-1}{}' \frac{1}{y_1-y_l} & \frac{1}{y_1-y_2} & \cdots & \frac{1}{y_1-y_{m-1}} \\ \frac{1}{y_2-y_0} & \frac{1}{y_2-y_1} & \sum_{l=0}^{m-1}{}' \frac{1}{y_2-y_l} & \cdots & \frac{1}{y_2-y_{m-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{y_{m-1}-y_0} & \frac{1}{y_{m-1}-y_1} & \frac{1}{y_{m-1}-y_2} & \cdots & \sum_{l=0}^{m-1}{}' \frac{1}{y_{m-1}-y_l} \end{pmatrix} \tag{2.68}$$

Clearly, $\operatorname{rank}\tilde{B}(\mathbf{y}) = \operatorname{rank} B(\mathbf{y})$. We prove that $\tilde{B}(\mathbf{y})$ has rank $m-1$ by explicitly

showing that $\tilde{B}(\mathbf{y})$ is similar to the Jordan block

$$J = \begin{pmatrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{2.69}$$

Proposition 2.6. *For $m \geq 2$, $\tilde{B}(\mathbf{y})$ has the Jordan decomposition*

$$\tilde{B}(\mathbf{y}) = P(\mathbf{y}) J P(\mathbf{y})^{-1}, \tag{2.70}$$

where

$$P(\mathbf{y}) = \begin{pmatrix} b_0(\mathbf{y}) & b_0(\mathbf{y})(y_0 - y_{m-1}) & \cdots & b_0(\mathbf{y})\frac{(y_0-y_{m-1})^{m-1}}{(m-1)!} \\ b_1(\mathbf{y}) & b_1(\mathbf{y})(y_1 - y_{m-1}) & \cdots & b_1(\mathbf{y})\frac{(y_1-y_{m-1})^{m-1}}{(m-1)!} \\ b_2(\mathbf{y}) & b_2(\mathbf{y})(y_2 - y_{m-1}) & \cdots & b_2(\mathbf{y})\frac{(y_2-y_{m-1})^{m-1}}{(m-1)!} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m-2}(\mathbf{y}) & b_{m-2}(\mathbf{y})(y_{m-2} - y_{m-1}) & \cdots & b_{m-2}(\mathbf{y})\frac{(y_{m-2}-y_{m-1})^{m-1}}{(m-1)!} \\ b_{m-1}(\mathbf{y}) & 0 & \cdots & 0 \end{pmatrix}. \tag{2.71}$$

Here the $b_j(\mathbf{y})$'s are defined as in (2.20).

Proof. The matrix $P(\mathbf{y})$ is of the form $D_1 V D_2$, where D_1, D_2 are invertible diagonal matrices and V is a Vandermonde matrix. Hence, $P(\mathbf{y})$ is invertible and the proof of (2.70) is equivalent to showing that $\tilde{B}(\mathbf{y})P(\mathbf{y}) = P(\mathbf{y})J$, that is, for $0 \leq j, n \leq m-1$,

$$\sum_{\substack{k=0\\k\neq j}}^{m-1} \frac{b_k(\mathbf{y})}{y_j - y_k} \frac{(y_k - y_{m-1})^n}{n!} + \sum_{\substack{l=0\\l\neq j}}^{m-1} \frac{b_j(\mathbf{y})}{y_j - y_l} \frac{(y_j - y_{m-1})^n}{n!} = b_j(\mathbf{y}) \frac{(y_j - y_{m-1})^{n-1}}{(n-1)!}, \tag{2.72}$$

where $\frac{1}{(-1)!} = 0$.

The proof is based on the counterclockwise integral defined for all $t \in \mathbb{C}$

$$J_{m,n}(t) = \frac{1}{2\pi i} \oint_{\Gamma} \frac{(z-t)^n}{\prod_{i=0}^{m-1} (z - y_i)} dz, \qquad 0 \leq n \leq m-1 \tag{2.73}$$

over a circle Γ of radius R large enough that it encloses all y_j's. Letting $R \to \infty$, we see that $J_{m,n} = \delta^{(0)}_{n-(m-1)}$ independent of t. On the other hand, note that the residue at y_k is $(-1)^{m-1} b_k(\mathbf{y})(y_k - t)^n$. Hence

$$\delta^{(0)}_{n-(m-1)} = J_{m,n} = (-1)^{m-1} \sum_{k=0}^{m-1} b_k(\mathbf{y})(y_k - t)^n. \tag{2.74}$$

Now

$$\frac{\partial b_k}{\partial y_j} = \begin{cases} \frac{b_k(\mathbf{y})}{y_k - y_j} & \text{for } j \neq k \\ \sum_{\substack{l=0\\l\neq j}}^{m-1} \frac{b_j(\mathbf{y})}{y_l - y_j} & \text{for } j = k \end{cases} \tag{2.75}$$

Hence, differentiating (2.74) with respect to y_j, leads to the identity

$$\sum_{\substack{k=0\\k\neq j}}^{m-1} \frac{b_k(\mathbf{y})}{y_k - y_j} (y_k - t)^n + \sum_{\substack{l=0\\l\neq j}}^{m-1} \frac{b_j(\mathbf{y})}{y_l - y_j} (y_l - t)^n + b_l(\mathbf{y}) n (y_j - t)^{n-1} = 0, . \tag{2.76}$$

Letting $t \to y_{m-1}$, one obtains (2.72). □

2.5 Asymptotics for the relaxed and the discrete minimization problem

In the following proposition, we evaluate the dependence on m of the solution $\mathbf{x} = \mathbf{x}^{(m)}$ of the relaxed minimization problem. For any fixed j, we show that $x_j^{(m)}$ converges as $m \to \infty$, and we compute the limit.

Proposition 2.7. *(a) For $K = K(m, \gamma)$ as in (2.60)*

$$\frac{2(m-1)^2}{(\cosh^{-1}\gamma)^2} - 1 \leq K \leq \frac{2m^2}{(\cosh^{-1}\gamma)^2} \tag{2.77}$$

(b) Set $\sigma := \frac{\pi^2}{\left(\cosh^{-1}(\gamma)\right)^2}$. Then for all m and all $1 \leq j \leq m-1$,

$$x_j^{(m)} \leq 1 + \sigma j^2 \tag{2.78}$$

(c) For any fixed $j \geq 1$,

$$\lim_{m\to\infty} x_j^{(m)} = 1 + \sigma j^2 \tag{2.79}$$

(d)

$$\lim_{m\to\infty} \frac{\left(\eta(\mathbf{x}^{(m)})\right)^{1/m}}{m^2} = \frac{1}{\cosh^{-1}\gamma} \tag{2.80}$$

Proof. We first provide bounds on β defined in (2.43). For a lower bound, write

$$\gamma = \frac{\cosh(2m-1)\beta)}{\cosh\beta} = \cosh(2m\beta) - \sinh(2m\beta)\tanh\beta \leq \cosh(2m\beta). \tag{2.81}$$

For an upper bound, we have

$$\gamma = \frac{\cosh(2m-1)\beta)}{\cosh\beta} = \cosh((2m-2)\beta) + \sinh((2m-2)\beta)\tanh\beta$$
$$\geq \cosh((2m-2)\beta). \tag{2.82}$$

We obtain the bounds

$$\frac{1}{2m}\cosh^{-1}\gamma \leq \beta \leq \frac{1}{2m-2}\cosh^{-1}\gamma. \tag{2.83}$$

This implies the upper bound for K

$$K = \frac{1}{2\sinh^2\beta} \leq \frac{1}{2\beta^2} \leq \frac{2m^2}{(\cosh^{-1}\gamma)^2}. \tag{2.84}$$

For the lower bound on K, we have by an elementary estimate

$$K = \frac{1}{2\sinh^2\beta} \geq \frac{1}{2\beta^2} - 1 \geq \frac{2(m-1)^2}{(\cosh^{-1}\gamma)^2} - 1. \tag{2.85}$$

This proves (a).

Also

$$x_j^{(m)} = 1 + 2K\sin^2\left(\frac{j\pi}{2m}\right) \leq 1 + \frac{4m^2}{(\cosh^{-1}\gamma)^2}\left(\frac{j\pi}{2m}\right)^2 = 1 + \sigma j^2, \tag{2.86}$$

which proves (b).

From (2.77)

$$\lim_{m\to\infty}\frac{K(m,\gamma)}{m^2} = \frac{2}{(\cosh^{-1}\gamma)^2}, \tag{2.87}$$

and so

$$\lim_{m\to\infty} x_j^{(m)} = 1 + \frac{2}{(\cosh^{-1}\gamma)^2} \lim_{m\to\infty} 2m^2 \sin^2\left(\frac{j\pi}{2m}\right) = 1 + \sigma j^2, \tag{2.88}$$

which proves (c).

Finally, from (2.83) we see that

$$\lim_{m\to\infty} 2m\beta(m,\gamma) = \cosh^{-1}\gamma, \tag{2.89}$$

and hence from (2.42)

$$\begin{aligned}\lim_{m\to\infty} \frac{(\eta(\mathbf{x}))^{1/m}}{m^2} &= \lim_{m\to\infty} \frac{1}{m^2}\left(\frac{\sinh(2m\beta)}{(2\sinh\beta)^{2m-1}\cosh\beta}\right)^{1/m} \\ &= \lim_{m\to\infty}\left(\frac{1}{4m^2\sinh^2\beta}\right)\left(\frac{2\sin\beta\sinh(2m\beta)}{\cosh\beta}\right)^{1/m} \\ &= \lim_{m\to\infty} \frac{K(m,\gamma)}{2m^2} = \frac{1}{(\cosh^{-1}\gamma)^2}, \end{aligned} \tag{2.90}$$

which proves (d). This completes the proof of the Proposition. □

The above results for the relaxed minimization problem allow us to draw conclusions for our original problem with the constraint that the filter locations $n_j^{(m)}$ are all integers.

With $n_1^{(m)} \equiv x_0^{(m)} \equiv 1$, we seek an integer sequence $\mathbf{n}^{(m)} = (n_2^{(m)}, \ldots, n_m^{(m)})$ such that $\mathbf{n}^{(m)}$ is subordinate in the sense of Definition 2.2 to $\mathbf{x}^{(m)} := (1 + K(1 + z_j))_{j=1}^{m-1}$, the solution of the relaxed minimization problem (2.11), (2.12). By Corollary 2.2, $f(\mathbf{n}^{(m)}) \le f(\mathbf{x}^{(m)}) \le \gamma$, so $\mathbf{n}^{(m)}$ satisfies (2.10). Note that for the $n_j^{(m)}$'s we use the original index set $j = 1, \ldots, m$ of Section 2.1, while for the $x_j^{(m)}$'s we retain the labels $j = 0, \ldots m-1$. In this section, we work with

the specific integer sequence $\mathbf{n}^{(m)} = (n_2^{(m)}, \ldots, n_m^{(m)})$ defined recursively by

$$n_{j+1}^{(m)} = \left\lceil n_j^{(m)} \frac{x_j^{(m)}}{x_{j-1}^{(m)}} \right\rceil, \qquad j = 1, \ldots, m-1, \tag{2.91}$$

where $n_1^{(m)} \equiv x_0^{(m)} \equiv 1$ as above and $\lceil s \rceil$ denotes the smallest integer greater or equal to s. This sequence is minimal amongst all integer sequences subordinate to $\mathbf{x}^{(m)}$ in the sense that if $\mathbf{k} = (k_2, \ldots, k_m)$ is any integer sequence such that $1 \leq \frac{k_2}{x_1^{(m)}} \leq \cdots \leq \frac{k_m}{x_{m-1}^{(m)}}$, then $k_j \geq n_j^{(m)}$ for all $j = 2, \ldots, m$. Indeed one has $k_2 \geq x_1^{(m)}$, which implies $k_2 \geq \lceil x_1^{(m)} \rceil = n_2^{(m)}$, and assuming by induction $k_j \geq n_j^{(m)}$, one obtains

$$k_{j+1} = \lceil k_{j+1} \rceil \geq \left\lceil x_j^{(m)} \frac{k_j}{x_{j-1}^{(m)}} \right\rceil \geq \left\lceil x_j^{(m)} \frac{n_j^{(m)}}{x_{j-1}^{(m)}} \right\rceil = n_{j+1}^{(m)}. \tag{2.92}$$

Definition 2.3. A sequence of integer vectors $\mathbf{k}^{(m)} = (k_2^{(m)}, \ldots, k_m^{(m)})$, of increasing length $m-1$, $m = 2, 3 \ldots$, with $f(k^{(m)}) \leq \gamma$ is said to be *asymptotically optimal* if

$$\lim_{m \to \infty} \left(\frac{\eta\left(\mathbf{k}^{(m)}\right)}{\eta\left(\mathbf{x}^{(m)}\right)} \right)^{1/m} = 1, \tag{2.93}$$

where $\mathbf{x}^{(m)}$ is the solution of (2.11), (2.12) as above.

The relevance of this definition will become clear after Theorem 2.10 below. The following lemma will be used to assess the asymptotic optimality of $\mathbf{n}^{(m)}$.

Lemma 2.8. $\mathbf{w}^{(m)} = (w_1^{(m)}, \ldots, w_{m-1}^{(m)})$ *defined by*

$$w_j^{(m)} = 1 + \sigma j^2, \tag{2.94}$$

for σ as in Proposition 2.7 (b), is subordinate to $\mathbf{x}^{(m)}$.

Proof. By Definition 2.2, we need to show that for $0 \leq j \leq m-2$ and $w_m^{(0)} \equiv x_0^{(m)} \equiv 1$,

$$\frac{w_{j+1}^{(m)}}{x_{j+1}^{(m)}} \geq \frac{w_j^{(m)}}{x_j^{(m)}}, \tag{2.95}$$

that is

$$(1+\sigma(j+1)^2)\left(1+2K\sin^2\left(\frac{j\pi}{2m}\right)\right) \geq (1+\sigma j^2)\left(1+2K\sin^2\left(\frac{(j+1)\pi}{2m}\right)\right), \tag{2.96}$$

or equivalently

$$\begin{aligned}&\left[\sigma(2j+1)-2K\left(\sin^2\left(\frac{(j+1)\pi}{2m}\right)-\sin^2\left(\frac{j\pi}{2m}\right)\right)\right]\\ &+\left[2\sigma K\left((j+1)^2\sin^2\left(\frac{j\pi}{2m}\right)-j^2\sin^2\left(\frac{(j+1)\pi}{2m}\right)\right)\right]\geq 0.\end{aligned} \tag{2.97}$$

We show that both these summands are nonnegative.

By Proposition 2.7 (a), $K \leq \frac{2m^2\sigma}{\pi^2}$, and so for the first summand, it is sufficient to show that

$$(2j+1)-\frac{4m^2}{\pi^2}\left(\sin^2\left(\frac{(j+1)\pi}{2m}\right)-\sin^2\left(\frac{j\pi}{2m}\right)\right)\geq 0. \tag{2.98}$$

Indeed, by standard trigonometric identities

$$\begin{aligned}\frac{4m^2}{\pi^2}\left(\sin^2\left(\frac{(j+1)\pi}{2m}\right)-\sin^2\left(\frac{j\pi}{2m}\right)\right) &= \frac{4m^2}{\pi^2}\sin\left(\frac{(2j+1)\pi}{2m}\right)\sin\left(\frac{\pi}{2m}\right)\\ &\leq 2j+1,\end{aligned} \tag{2.99}$$

which proves (2.98).

On the other hand, the positivity of the second summand follows from the fact that the function $\frac{\sin(y)}{y}$ is decreasing on $[0, \frac{\pi}{2}]$. This completes the proof of (2.97) and hence the proof of the Lemma. □

Theorem 2.9. *If σ, defined in Proposition 2.7, is an integer, then the sequence $\mathbf{n}^{(m)}$, $m = 2, 3, \ldots$, defined by (2.91), is both asymptotically optimal and subordinate to $\mathbf{x}^{(m)}$. If σ is not an integer, no sequence of integer vectors can have both these properties.*

Proof. If σ is not an integer, then $\lim_{m\to\infty} x_1^{(m)} = 1 + \sigma$ is not an integer, and so for any sequence $\mathbf{k}^{(m)} = (k_2^{(m)}, \ldots, k_m^{(m)})$, $m = 2, 3, \ldots$, of integer vectors subordinate to $\mathbf{x}^{(m)}$,

$$\limsup_{m\to\infty} \left(\frac{\eta(\mathbf{k})}{\eta(\mathbf{x})}\right)^{1/m} \geq \limsup_{m\to\infty} \frac{k_2^{(m)}}{x_1^{(m)}} \geq \lim_{m\to\infty} \frac{\lceil x_1^{(m)} \rceil}{x_1^{(m)}} = \frac{1 + \lceil \sigma \rceil}{1 + \sigma} > 1. \tag{2.100}$$

Hence $\mathbf{k}^{(m)}$ cannot be asymptotically optimal.

Now consider the case that σ is an integer. Then by Lemma 2.8, $\mathbf{w}^{(m)} = (w_1^{(m)}, \ldots, w_{m-1}^{(m)})$ given by

$$w_j^{(m)} = 1 + \sigma j^2 \tag{2.101}$$

is an integer sequence subordinate to $\mathbf{x}^{(m)}$. It follows that for $j = 1, \ldots, m-1$, one has $x_j^{(m)} \leq {n_{j+1}}^{(m)} \leq w_j^{(m)}$, as $\mathbf{n}^{(m)}$ is the minimal integer sequence subordinate to $\mathbf{x}^{(m)}$.

From Proposition 2.7 (a) we see that

$$\frac{2K}{m^2} \geq \frac{4s}{\pi^2}\left(1 - \frac{C_1}{m}\right) \tag{2.102}$$

for some constant $C_1 < \infty$. Together with the elementary fact that $\left(\frac{\sin x}{x}\right)^2 \geq$

$1 - C_2 x^2$ for a sufficiently large constant C_2, this implies that for $1 \leq j \leq m^{2/3}$ and some constant $C_3 < \infty$

$$\begin{aligned} x_j^{(m)} &= 1 + 2K \sin^2\left(\frac{j\pi}{2m}\right) \geq 1 + \sigma\left(1 - \frac{C_1}{m}\right)\left(j^2\left(1 - C_2\frac{\pi^2}{4}\frac{j^2}{m^2}\right)\right) \\ &\geq (1 + \sigma j^2)\left(1 - \frac{C_3}{m^{2/3}}\right). \end{aligned} \tag{2.103}$$

Then for $1 \leq j \leq m^{2/3}$ and some $C_4 < \infty$

$$\frac{{n_{j+1}}^{(m)}}{{x_j}^{(m)}} \leq \frac{{w_j}^{(m)}}{{x_j}^{(m)}} \leq \frac{1}{1 - \frac{C_3}{m^{2/3}}} \leq 1 + \frac{C_4}{m^{2/3}}. \tag{2.104}$$

Now

$$\frac{{n_{j+1}}^{(m)}}{{x_j}^{(m)}} = \frac{1}{{x_j}^{(m)}}\left\lceil n_j^{(m)}\frac{x_j^{(m)}}{x_{j-1}^{(m)}}\right\rceil \leq \frac{1}{{x_j}^{(m)}}\left(n_j^{(m)}\frac{x_j^{(m)}}{x_{j-1}^{(m)}} + 1\right) = \frac{n_j^{(m)}}{x_{j-1}^{(m)}} + \frac{1}{{x_j}^{(m)}}. \tag{2.105}$$

Combining (2.102) together with the elementary lower bound $\frac{\sin x}{x} \geq \frac{2}{\pi}$ for $0 \leq x \leq \frac{\pi}{2}$, we obtain $x_j^{(m)} \geq C_5 j^2$, $j \geq 1$, for some constant $C_5 > 0$. By repeated application of (2.105), one then obtains for $m^{2/3} < j \leq m - 1$ and some constant $C_6 < \infty$

$$\frac{n_{j+1}^{(m)}}{x_j^{(m)}} \leq \frac{n_j^{(m)}}{x_{j-1}^{(m)}} + \frac{1}{C_5 j^2} \leq \cdots \leq \frac{n_{\lfloor m^{2/3}\rfloor}^{(m)} + 1}{x_{\lfloor m^{2/3}\rfloor}^{(m)}} + \sum_{l=\lfloor m^{2/3}\rfloor+1}^{j} \frac{1}{C_5 l^2} \leq 1 + \frac{C_4}{m^{2/3}} + \frac{C_6}{m^{2/3}} \tag{2.106}$$

Thus there exists some constant $C_7 < \infty$, such that for all $1 \leq j \leq m - 1$,

$$\frac{{n_{j+1}}^{(m)}}{{x_j}^{(m)}} \leq 1 + \frac{C_7}{m^{2/3}}. \tag{2.107}$$

We conclude that

$$1 \leq \frac{\eta(\mathbf{n}^{(m)})}{\eta(\mathbf{x}^{(m)})} = \prod_{j=1}^{m-1} \frac{n_{j+1}^{(m)}}{x_j{}^{(m)}} \leq \left(1 + \frac{C_7}{m^{2/3}}\right)^m, \tag{2.108}$$

which implies that

$$\lim_{m\to\infty} \left(\frac{\eta(\mathbf{n}^{(m)})}{\eta(\mathbf{x}^{(m)})}\right)^{1/m} = 1, \tag{2.109}$$

and hence $\mathbf{n}^{(m)}$ is asymptotically optimal. □

Figure 2.1 illustrates the fact that for σ non-integer, the resulting sequence $\mathbf{n}^{(m)}$ is not asymptotically optimal. As an example, we consider the case $\gamma = 1.5$, corresponding to $\sigma \approx 10.66$. It turns out that for $m \geq 18$, one has $n_2 > w_1$, and so the above argument breaks down. The plot in Figure 2.1 compares, for the smallest such order $m = 18$, the case $\gamma = 1.5$ with the case corresponding to the (next larger) integer value $\sigma = 11$, i.e., $\gamma \approx 1.48$. The values of the n_j's arising in these two cases differ by at most 1, so they are indistinguishable in the plot. Hence, allowing for $\gamma = 1.5$ instead of $\gamma = 1.48$ leads to almost no reduction of $\eta(\mathbf{n})^{1/m}$, although it does lead to a significant reduction of $\eta(\mathbf{x})^{1/m}$. This corresponds to the fact that only in the latter case, one has asymptotic optimality.

Theorem 2.10. *For all $1 < \gamma < 2$ such that $\sigma = \frac{\pi^2}{\left(\cosh^{-1}\gamma\right)^2}$ is an integer, all $\Sigma\Delta$ modulators corresponding to filters $h^{(m)}$ minimally supported at positions $1, n_2^{(m)}, \ldots, n_m^{(m)}$ are stable for all input sequences y with $\|y\|_{\ell^\infty} \leq \mu = 2 - \gamma$. Furthermore, the family consisting of the $\Sigma\Delta$ modulators corresponding to the filters $\left\{h^{(m)}\right\}_{m=2}^{\infty}$ for all orders m gives rise to exponential error decay: For any rate constant $r < r_0 := \frac{\pi}{e^2\sigma \ln 2}$, there exists a constant $C = C(r)$ such that*

$$\|e_\lambda\|_{L^\infty} \leq C 2^{-r\lambda}. \tag{2.110}$$

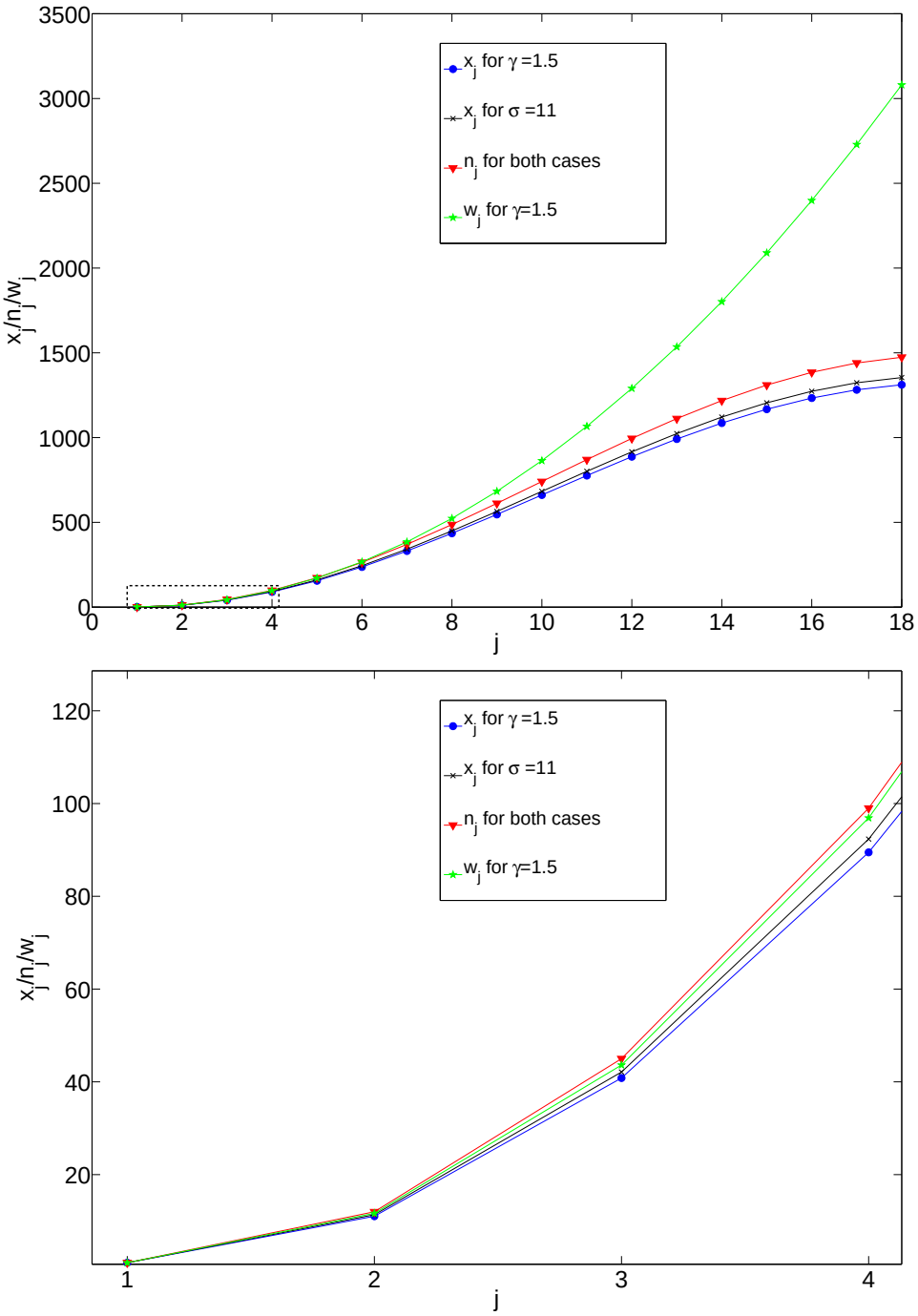

Figure 2.1: The x_j's and n_j's as a function of j for $m = 18$ and $\gamma = 1.5$ or $\sigma = 11$, respectively. For comparison, the w_j's for $\gamma = 1.5$ are included in the plot. The second plot enlarges the dashed box in the first plot, showing that when $\gamma = 1.5$ and hence σ is not an integer, one has $n_j > w_j$ for small j.

Proof. Stability follows from the fact that $\mathbf{n}^{(m)}$ is subordinate to $\mathbf{x}^{(m)}$, which satisfies the stability condition $f(\mathbf{x}^{(m)}) \leq \gamma$.

Choose the reconstruction kernel φ such that the corresponding λ_0 as in Theorem 1.1 satisfies $\lambda_0 < \sqrt{\frac{r_0}{r}}$. Now let $g^{(m)}$ be such that $\Delta^m g^{(m)} = \delta^{(0)} - h^{(m)}$, as in Definition 1.2. Then by Theorem 1.3, we have the error bound

$$\|e_\lambda\|_{L^\infty} \leq \|g^{(m)}\|_{\ell^1} \|v\|_{\ell^\infty} \|\varphi\|_{L^1} \pi^m \lambda_0^m \lambda^{-m}, \tag{2.111}$$

where v solves (1.50).

Recall that our construction yields $\|v\|_{\ell^\infty} \leq 1$. Furthermore, by Theorem 2.9 and Proposition 2.7 (d), we have that

$$\lim_{m\to\infty} \frac{\left(\eta(\mathbf{n}^{(m)})\right)^{1/m}}{m^2} = \frac{1}{(\cosh^{-1}\gamma)^2} \tag{2.112}$$

and hence by (2.4)

$$\|g^{(m)}\|_{\ell^1} = \frac{\eta(\mathbf{n}^{(m)})}{m!} = \left(\frac{e}{(\cosh^{-1}\gamma)^2}\right)^m m^m \left(1 + o(1)\right)^m. \tag{2.113}$$

Now consider $m \geq M(r)$ large enough to ensure that the $(1 + o(1))$-factor is less than $\sqrt{\frac{r_0}{r}}$. Then

$$\|e_\lambda\|_{L^\infty} \leq \|\varphi\|_{L^1} \left(\frac{\pi e}{(\cosh^{-1}\gamma)^2}\right)^m m^m \left(\frac{r_0}{r}\right)^m \lambda^{-m} = \|\varphi\|_{L^1} \left(\frac{e\sigma}{\pi}\right)^m m^m \left(\frac{r_0}{r}\right)^m \lambda^{-m}. \tag{2.114}$$

Now as explained in section 1.2.3, we choose, for each λ, the filter $h^{(m)}$ that leads to the minimal error bound. Up to a constant, the bound given in Equation (2.114) is of the form $m^m \alpha^{-m}$ for some $\alpha = \alpha(\lambda) \in \mathbb{R}$. Now in [7] it is shown that there

exists a uniform constant C' such that for all $\alpha \in \mathbb{R}$

$$\min_{m \in \mathbb{N}} \; m^m \alpha^{-m} \leq C' e^{-\alpha/e}. \tag{2.115}$$

Minimizing not over all $m \in \mathbb{N}$, but only over $m \geq M(r)$ introduces an additional constant, but for some constant $C = C(r)$ we still obtain

$$\|e_\lambda\|_{L^\infty} \leq \|\varphi\|_{L^1} \min_{m \geq M(r)} \left(\frac{e\sigma}{\pi}\right)^m m^m \left(\frac{r_0}{r}\right)^m \lambda^{-m} \tag{2.116}$$

$$\leq C \exp\left(-\frac{\pi}{e^2 \sigma} \frac{r}{r_0} \lambda\right) = C 2^{-r\lambda}, \tag{2.117}$$

which proves the theorem. □

Remark: The smallest integer σ such that the stability constraint $\|h\|_{\ell^1} \leq \gamma$ is satisfied for some $\gamma < 2$ is $\sigma = 6$. In this case, Theorem 2.10 yields exponential error decay for rate constants $r < r_0 \approx 0.102$. This is the fastest error decay currently known to be achievable for $\Sigma\Delta$ modulation. The previously best known bound for the achievable rate constant was $r_0 \approx 0.088$ [11].

Chapter 3

Stability analysis for MCR $\Sigma\Delta$ modulators

3.1 MCR modulators

The constructions presented in Chapter 2 employ filters h with finite impulse response together with the greedy quantization rule. These filters, however, have a large number of filter tabs. Indeed, Güntürk showed in [7] that the circuit complexity of $\Sigma\Delta$ modulators based on FIR filters which satisfy Stability Criterion (1.65) grows quadratically in the order m. More precisely, there exists a constant C such that every FIR filter $h = \delta^{(0)} - \Delta^m g$ (for some $g \in \ell^1$) satisfying Criterion (1.65) has at least Cm^2 filter tabs. Filters with many tabs result in higher hardware cost and require a long sequence of delays which can lead to impaired accuracy. For these reasons, engineers prefer to use filters with infinite impulse response that do not require long delay sequences [16]. Ideally, the variables in an m-th order $\Sigma\Delta$ modulator would be stored only for m time instances. For the corresponding

generating function $H(z) = \frac{B(z)}{A(z)}$, this amounts to choosing both $A(z)$ and $B(z)$ to be polynomials of degree m. The resulting circuits are of minimal complexity with rational transfer functions; we use the acronym MCR for these schemes.

As h defines an m-th order modulator, $1 - H(z) = \frac{A(z)-B(z)}{A(z)}$ should be divisible by $(1-z)^m$, which, together with the facts $a_0 = 0$ and $b_0 = 1$, imply that $A(z) - B(z) = (1-z)^m$ and $1 - H(z) = \frac{(1-z)^m}{A(z)}$. Once again, we can write $1 - h = \Delta^m g$, where, up to a shift, g has the generating function $G = \frac{1}{A}$. Thus $G(z)A(z) = 1$ and, again up to a shift, $a * g = \delta^{(0)}$ for the vector a that has A as its generating function. We call g the *convolutional inverse* of a and write $g = a^{-1}$. While the convolutional inverse a^{-1} can always be defined through the power series expansion of $\frac{1}{A(z)}$, nothing is known, a priori, about the decay properties or even boundedness of the entries of such sequences. The following lemma provides a criterion to ensure that $g = a^{-1}$ and the corresponding h are in ℓ^1.

Lemma 3.1. *If a sequence a has the generating function*

$$A(z) = \prod_{j=1}^{m} (1 - \xi_j z) \tag{3.1}$$

for some $\xi_1, \ldots, \xi_m \in \mathbb{C}$ with $|\xi_j| < 1$ for all j, then a has a convolutional inverse a^{-1} in ℓ^1.

Proof. Define

$$S^{(i)}(z) := \frac{1}{1 - \xi_i z} = \sum_{l=0}^{\infty} \xi_i^l z^l. \tag{3.2}$$

Then $S^{(i)}$ is the generating function of the sequence $s^{(i)} \in \ell^1$ given by $s_l^{(j)} = \xi_j^l$. Now the ℓ^1-sequence $s = s^{(1)} * \cdots * s^{(m)}$ corresponds to the generating function $S = \prod_{j=1}^{m} S^{(i)} = \frac{1}{A}$. We conclude that since $s * a$ has the generating function

$S(z)A(z) = 1$, one has $s * a = \delta^{(0)}$ and $s \in \ell^1$ is the convolutional inverse of a. $\square$

We are now ready to make a precise definition for a MCR $\Sigma\Delta$ modulator:

Definition 3.1. We say that the $\Sigma\Delta$ modulator given by the difference equation

$$v_n = h * v + y_n - q_n \tag{3.3}$$

together with the greedy quantization rule

$$q_n = sign(h * v + y_n) \tag{3.4}$$

is an *MCR* $\Sigma\Delta$ *modulator* (or just *MCR modulator*) if

$$\delta^{(0)} - h = \Delta^m a^{-1} \tag{3.5}$$

for a finite vector a with generating function $A(z)$ given by

$$A(z) = \prod_{j=1}^{m} (1 - \xi_j z) \tag{3.6}$$

for some $\xi_1, \ldots, \xi_m \in \mathbb{C}$ such that $|\xi_j| < 1$ for all $1 \leq j \leq m$.

3.2 A criterion for the roots of the denominator

Even though MCR modulators and other modulators that employ similar infinite impulse response filters are commonly used in practice, there is hardly any rigorous stability analysis available for such quantization schemes. In this section, we show using an explicit construction that stable MCR $\Sigma\Delta$ modulators exist for

all orders. We will work with the stability criterion given in theorem 1.6, i.e., we seek filters h with small ℓ^1-norm. In the framework of MCR modulators, this amounts to optimizing the location of the ξ_j introduced in Definition 3.1. While engineers observe the best error decay when the ξ_i appear in complex conjugate pairs (compare [16]), it turns out that the scenario where all the roots are chosen to be real can be generalized to arbitrary orders more easily.

Furthermore, it will be convenient to work with a different set of parameters. Instead of $\{\xi_1, \ldots, \xi_m\}$, we will use the parameters $\{r_1, \ldots, r_{m-1}, \xi_m\}$, where, for $j = 1, \ldots, m-1$, $r_j := \frac{1-\xi_j}{1-\xi_{j+1}}$ and ξ_m is unchanged. Indeed, for $k = 1, \ldots, m-1$, one has

$$\xi_k = 1 - (1 - \xi_m) \prod_{j=k}^{m-1} r_j, \tag{3.7}$$

and so $\{r_1, \ldots r_{m-1}, \xi_m\}$ is an equivalent set of independent parameters.

Proposition 3.2. *For $0 < r_j < 1$, $j = 1, \ldots, m-1$ and $0 \leq \xi_m < 1$, let a be the finite vector that has the generating function $A(z) = \prod_{j=1}^{m}(1-\xi_j z)$. Here, for $j < m$, the ξ_j's and the r_j's are related as in Equation (3.7). Then for $h := \delta^{(0)} - \Delta^m a^{-1}$ as above, one has*

$$\|h\|_{\ell^1} \leq 1 + \sum_{j=1}^{m-1} \frac{2^{m-j}}{\frac{1}{r_j} - 1} \tag{3.8}$$

Proof. The generating function of $\delta^{(0)} - h$ is given by

$$1 - H(z) = \frac{(1-z)^m}{\prod_{j=1}^{m}(1 - \xi_j z)} \tag{3.9}$$

Thus we can write

$$\delta^{(0)} - h = \eta^{(\xi_1)} * \cdots * \eta^{(\xi_m)}, \tag{3.10}$$

where

$$\eta^{(\xi_j)} = \Delta s^{(j)} = \Delta(1, \xi_j, \xi_j^2, \dots) \tag{3.11}$$

corresponds, up to a shift, to the generating function

$$E^{(\xi_j)} = \frac{1-z}{1-\xi_j z}. \tag{3.12}$$

Equivalently,

$$\begin{aligned} h &= \delta^{(0)} - \eta^{(\xi_1)} * \cdots * \eta^{(\xi_m)} \\ &= (\delta^{(0)} - \eta^{(\xi_1)}) * \eta^{(\xi_2)} * \cdots * \eta^{(\xi_m)} + \delta^{(0)} - \eta^{(\xi_2)} * \cdots * \eta^{(\xi_m)} \\ &= \dots \\ &= \sum_{j=1}^{m} (\delta^{(0)} - \eta^{(\xi_j)}) * \eta^{(\xi_{j+1})} * \cdots * \eta^{(\xi_m)}. \end{aligned} \tag{3.13}$$

From the identity

$$1 - E^{(\xi_j)} = \frac{(1-\xi_j z) - (1-z)}{1-\xi_j z} = \frac{z(1-\xi_j)}{1-\xi_j z}, \tag{3.14}$$

we obtain that up to a shift

$$\delta^{(0)} - \eta^{(\xi_j)} = (1-\xi_j)(0, 1, \xi_j, \dots). \tag{3.15}$$

In particular, all entries of $\delta^{(0)} - \eta^{(\xi_j)}$ are non-negative, and we see that

$$\|\delta^{(0)} - \eta^{(\xi_j)}\|_{\ell^1} = (1-\xi_j) \sum_{l=0}^{m} \xi_j^l = 1 \tag{3.16}$$

and, as $\eta_1^{(\xi_j)} = 1$, $\|\eta^{(\xi_j)}\|_{\ell^1} = 2$.

Furthermore, for $j < m$,

$$\begin{aligned}\left(1 - E^{(\xi_j)}(z)\right) E^{(\xi_{j+1})} &= \frac{z(1-\xi_j)}{1-\xi_j z}\frac{1-z}{1-\xi_{j+1}z} \\ &= (1-z)\frac{(1-\xi_j)}{\xi_j - \xi_{j+1}}\left[\frac{1}{1-\xi_j z} - \frac{1}{1-\xi_{j+1}z}\right], \end{aligned} \tag{3.17}$$

and correspondingly for the $\eta^{(\xi_j)}$'s

$$(\delta^{(0)} - \eta^{(\xi_j)}) * \eta^{(\xi_{j+1})} = \frac{(1-\xi_j)}{\xi_j - \xi_{j+1}}\Delta\left(0, \xi_j - \xi_{j+1}, \xi_j^2 - \xi_{j+1}^2, \dots\right). \tag{3.18}$$

To find the ℓ_1-norm of $\sigma^{(j)} := \Delta\left(0, \xi_j - \xi_{j+1}, \xi_j^2 - \xi_{j+1}^2, \dots\right)$, we note that there is a unique integer N such that

$$\sigma_n^{(j)} \geq 0 \text{ for } n \leq N$$
$$\sigma_n^{(j)} < 0 \text{ for } n > N.$$

Indeed,

$$\begin{aligned} &\sigma_n^{(j)} = [\Delta\left(0, \xi_j - \xi_{j+1}, \xi_j^2 - \xi_{j+1}^2, \dots\right)]_n < 0 \\ \Leftrightarrow &-\xi_j^{n-1}(1-\xi_j) + \xi_{j+1}^{n-1}(1-\xi_{j+1}) < 0 \\ \Leftrightarrow &n > \frac{\ln(1-\xi_{j+1}) - \ln(1-\xi_j)}{\ln(\xi_j) - \ln(\xi_{j+1})} + 1 \\ \Rightarrow &N = \left\lfloor \frac{\ln(1-\xi_{j+1}) - \ln(1-\xi_j)}{\ln(\xi_j) - \ln(\xi_{j+1})} \right\rfloor + 1. \end{aligned} \tag{3.19}$$

Hence we calculate:

$$\begin{aligned}\left\|\sigma^{(j)}\right\|_{\ell^1} &= \sum_{n=0}^{N} \sigma_n^{(j)} - \sum_{n=N+1}^{\infty} \sigma_n^{(j)} \\ &= \sum_{n=0}^{N} [\Delta\left(0, \xi_j - \xi_{j+1}, \dots\right)]_n - \sum_{n=N+1}^{\infty} [\Delta\left(0, \xi_j - \xi_{j+1}, \dots\right)]_n \qquad (3.20) \\ &= 2\xi_j^N - 2\xi_{j+1}^N < 2 \qquad (3.21)\end{aligned}$$

For the last equality, we used that both sums in Equation (3.20) are telescoping.

We conclude that

$$\|(\delta^{(0)} - \eta^{(\xi_j)}) * \eta^{(\xi_{j+1})}\|_{\ell^1} < 2\frac{(1-\xi_j)}{\xi_j - \xi_{j+1}} \qquad (3.22)$$

and

$$\begin{aligned}\|h\|_{\ell^1} &\leq \sum_{j=1}^{m} \|(\delta^{(0)} - \eta^{(\xi_j)}) * \eta^{(\xi_{j+1})} * \cdots * \eta^{(\xi_m)}\|_{\ell^1} \\ &\leq \sum_{j=1}^{m-1} \|(\delta^{(0)} - \eta^{(\xi_j)}) * \eta^{(\xi_{j+1})}\|_{\ell_1} \|\eta^{(\xi_{j+2})}\|_{\ell^1} \cdots \|\eta^{(\xi_m)}\|_{\ell^1} + \|\delta^{(0)} - \eta^{(\xi_m)}\|_{\ell^1} \\ &< \sum_{j=1}^{m-1} 2\frac{(1-\xi_j)}{\xi_j - \xi_{j+1}} 2^{m-j-1} + 1. \\ &= 1 + \sum_{j=1}^{m-1} \frac{2^{m-j}}{\frac{1}{r_j} - 1}. \qquad (3.23)\end{aligned}$$

□

Corollary 3.3. *Under the same assumptions as in Proposition 3.2, one has*

$$\lim_{r_1, \dots, r_{m-1} \to 0} \|h\|_{\ell^1} = 1. \qquad (3.24)$$

Proof. The result follows directly from (3.8) combined with the fact that $\|h\|_{\ell_1} \geq 1$, as noted in section 1.2.4. □

Remark: In terms of the original parameters, letting all $r_j \to 0$ amounts to letting all ξ_j except possibly ξ_m tend to 1 so that each ξ_j converges faster than ξ_{j+1}.

Theorem 3.4. *For each $\mu < 1$, there exists an infinite family of MCR $\Sigma\Delta$ modulators that has the following properties:*

- *All the corresponding filters h satisfy $\|h\|_{\ell^1} \leq 2 - \mu$.*
- *The error decay in λ that results from choosing, for each λ, the optimal modulator from the family can be bounded by*

$$\|e_\lambda\|_{\ell^\infty} \leq C\lambda^{-\kappa\sqrt{\log \lambda}} \tag{3.25}$$

for some constants $C < \infty$, $\kappa > 0$.

Proof. Fix $\mu < 1$. It follows from Proposition 3.3, that for each m, choosing the r_j's, $j = 1, \ldots, m-1$, small enough together with an arbitrary $0 \leq \xi_m < 1$ yields an MCR $\Sigma\Delta$ modulator $\mathcal{M}^{(m)}$ such that the corresponding filter h satisfies $\|h\|_{\ell^1} \leq 2 - \mu$. To estimate the error decay that the infinite family of modulators $\{\mathcal{M}^{(m)}\}_{m=1}^{\infty}$ yields, we bound

$$\|g\|_{\ell^1} = \|s^{(1)} * \cdots * s^{(m)}\|_{\ell^1} \leq \prod_{j=1}^{m} \|s^{(j)}\|_{\ell^1} = \prod_{j=1}^{m} \frac{1}{1-\xi_j}. \tag{3.26}$$

With Relation (3.7), this implies

$$\|g\|_{\ell^1} \leq \prod_{j=1}^{m-1} r_j^{-j} \left(\frac{1}{1-\xi_m}\right)^m =: g^*. \tag{3.27}$$

Our goal is now to minimize g^* while ensuring that $\|h\|_{\ell^1} \leq 2-\mu$. A sufficient condition for this constraint to hold is that the upper bound given in Proposition 3.2 is less or equal to $2-\mu$. In this case, each denominator in Equation (3.8) is greater or equal to $\frac{2}{1-\mu}$, or equivalently one has $\frac{1}{r_j} - 1 \geq \gamma \frac{1}{r_j}$ for all $1 \leq j \leq m-1$, where $\gamma := \frac{2}{3-\mu}$. Then

$$\|h\|_{\ell^1} \leq 1 + \sum_{j=1}^{m-1} \frac{2^{m-j}}{\frac{1}{r_j} - 1} \leq 1 + \sum_{j=1}^{m-1} \frac{2^{m-j}}{\gamma} r_j =: h^*. \tag{3.28}$$

We will now minimize g^* subject to the $h^* \leq 2-\mu$. From Equations (3.28) and (3.27), one sees that leaving all r_j fixed and letting ξ_m decrease causes g^* to decrease while keeping the value of h^* fixed. Hence, one should set ξ_m to the minimal admissible value $\xi_m = 0$ and consider the following minimization problem:

$$\begin{aligned} &\text{Minimize } g^* = \prod_{j=1}^{m-1} r_j^{-j} \\ &\text{over } \{(r_1, \ldots, r_{m-1}) \in (0,1)^{m-1} : h^* = 1 + \sum_{j=1}^{m-1} \frac{2^{m-j}}{\gamma} r_j \leq 2-\mu\}. \end{aligned} \tag{3.29}$$

We will now set up the associated Lagrange multiplier problem. Similarly to the discussion in Chapter 2, one can show the existence of a minimizer $\mathbf{r}_{min}$ that does not lie on the boundary and that yields equality in the constraint, $h^*(\mathbf{r}_{min}) = 2-\mu$. In contrast to the minimization problem discussed in that chapter, however, Minimization Problem (3.29) does not describe the underlying problem precisely

but is meant to heuristically determine some filter with good error decay. For this reason, it is not important to show that the critical point obtained from solving the Lagrange multiplier equations under the equality constraint $h^* = 2-\mu$ is really the absolute minimum, and we leave the details to the reader. To set up the Lagrange multiplier equation, calculate for $1 \leq j \leq m-1$

$$\frac{\partial}{\partial r_j} g^* = -\frac{j g^*}{r_j} \tag{3.30}$$

and

$$\frac{\partial}{\partial r_j} h^* = -\frac{2^{m-j}}{\gamma}. \tag{3.31}$$

Hence, the Lagrange multiplier formulation of Minimization Problem (3.29) reads as follows:

For every critical point $(r_1, r_2, \ldots, r_{m-1})$ there exists $\zeta \in \mathbb{R}$ such that for all $j = 1, \ldots, m-1$ one has

$$\frac{j g^*}{r_j} = \zeta \frac{2^{m-j}}{\gamma}. \tag{3.32}$$

Setting $\tilde{\zeta} = \frac{2^m \zeta}{g^*}$, we obtain

$$r_j = \frac{\gamma j 2^j}{\tilde{\zeta}}. \tag{3.33}$$

Substituting these values of r_j into the constraint, we obtain

$$2 - \mu = h^* = 1 + \sum_{j=1}^{m-1} \frac{2^{m-j}}{\gamma} r_j = 1 + \sum_{j=1}^{m-1} \frac{j 2^m}{\tilde{\zeta}} = 1 + \frac{m(m-1)2^m}{\tilde{\zeta}}, \tag{3.34}$$

and consequently

$$\tilde{\zeta} = \frac{m(m-1)2^m}{1-\mu} \tag{3.35}$$

as well as

$$r_j = \frac{\gamma j(1-\mu)}{m(m-1)2^{m-j}}. \tag{3.36}$$

Using the identity $\sum_{j=1}^{m-1} j(m-j) = \frac{1}{6}m(m-1)(m+1)$, this yields

$$\begin{aligned} g^* &= \prod_{j=1}^{m-1} r_j^{-j} \\ &= \prod_{j=1}^{m-1} \left(\frac{m(m-1)2^{m-j}}{\gamma j(1-\mu)} \right)^j \\ &= \left(\frac{m(m-1)}{\gamma(1-\mu)} \right)^{\frac{1}{2}m(m-1)} \frac{2^{\frac{1}{6}m(m-1)(m+1)}}{\prod_{j=1}^{m-1} j^j}. \end{aligned} \tag{3.37}$$

Hence, by Theorem 1.3 and using the fact that $\|v\|_{\ell^\infty} \leq 1$, we bound:

$$\begin{aligned} \|e_\lambda\|_{L^\infty} &\leq \|g\|_{\ell^1} \|\varphi\|_{L^1} \pi^m \lambda^{-m} \\ &\leq g^* \|\varphi\|_{L^1} \pi^m \lambda^{-m} \\ &= \frac{(m(m-1))^{\frac{1}{2}m(m-1)}\, 2^{\frac{1}{6}m(m-1)(m+1)}}{\gamma \prod_{j=1}^{m-1} j^j} \pi^m \lambda^{-m} \end{aligned} \tag{3.38}$$

Now for any $\epsilon > 0$, $\lim_{m\to\infty} 2^{-\epsilon m^3} \frac{(m(m-1))^{\frac{1}{2}m(m-1)}}{\gamma \prod_{j=1}^{m-1} j^j} \pi^m = 0$, so the term $2^{\frac{1}{6}m^3}$ dominates and for any $\alpha > \frac{1}{6}$, we can find a constant $C < \infty$ such that for all m

$$\|e_\lambda\|_{L^\infty} \leq C 2^{\alpha m^3} \lambda^{-m}. \tag{3.39}$$

For each λ, minimize the expression on the right hand side over m, allowing for

arbitrary real values of m. The minimizer can be easily computed to be

$$m_{min} = \sqrt{\frac{1}{3\alpha} \log_2 \lambda} \tag{3.40}$$

yielding the minimum

$$2^{\alpha m_{min}^3} \lambda^{-m_{min}} = \lambda^{-\frac{2}{3}\sqrt{\frac{1}{3\alpha} \log_2 \lambda}}. \tag{3.41}$$

Now m is constraint to be an integer, but the correction term that arises from choosing $\lfloor m_{min} \rfloor$ instead of m_{min} is also dominated and can be absorbed into the constant C (for a similar argument with all of the details see [7]). We conclude that

$$\|e_\lambda\|_{L^\infty} \leq C\lambda^{-\frac{2}{3}\sqrt{\frac{1}{3\alpha} \log_2 \lambda}}, \tag{3.42}$$

which proves the proposition. □

Remark: While the error decay bound established by Theorem 3.4 is faster than any inverse polynomial, it is considerably slower than the exponential error decay bound established in chapter 2 or even the subexponential error decay established in [4] for schemes in canonical form with non-standard quantizers. The value of the schemes considered in this chapter lies mostly in the low complexity of the underlying circuit architecture. This observation is in line with observations in the engineering literature (see for example [16]) stating that the stability criterion given in Theorem 1.6 is "conservative" in the sense that it is too strict to allow for good error decay.

3.3 Application to schemes with linear quantization rules

For an m-th order MCR $\Sigma\Delta$ modulator, the quantizer can be expressed explicitly in terms of the canonical variables $u_n = (g * v)_n$ as introduced in section 1.2.2. Indeed, as $g = a^{-1}$ and hence $h = \delta^{(0)} - \Delta^m g = a * g - \Delta^m g$, we obtain

$$\begin{aligned} q_n &= sign([h * v]_n + y_n) \\ &= sign([(a - \Delta^m) * g * v]_n + y_n) \\ &= sign([(a - \Delta^m) * u]_n + y_n). \end{aligned} \tag{3.43}$$

So if a particular choice of a yields a stable MCR $\Sigma\Delta$ modulator, it follows that for $d = a - \Delta^m$, the $\Sigma\Delta$ modulator in canonical form

$$\Delta^m u_n = y_n - q_n \tag{3.44}$$

with the *linear quantization rule*

$$q_n = sign([d * u]_n + y_n) \tag{3.45}$$

is also stable.

In this manner, every statement about the stability of MCR modulators is equivalent to a statement about the stability of a corresponding scheme in its canonical variables with a linear quantization rule. For example, Theorem 3.4 also proves the existence of stable schemes in canonical form with a linear quantization rule for all orders.

Chapter 4

A novel stability criterion

In Chapters 2 and 3, we have shown stability for two families of $\Sigma\Delta$ modulators by applying the stability criterion $\|h\|_{\ell^1} \leq 2 - \mu$ in conjunction with the greedy quantization rule. In contrast, in Section 4.1, we consider modulators with quantization rules different from the greedy rule. We show that the stability criterion is robust with respect to small ℓ^1-perturbations of the coefficients used for the quantizer input. In the remainder of this chapter, we apply these results to derive a generalized stability criterion for the greedy rule.

4.1 A stability criterion for approximately greedy quantization rules

Theorem 4.1. *For $h, \tilde{h} \in \ell^1$ causal sequences, consider the $\Sigma\Delta$ modulator given by*

$$v_n = (h * v)_n + y_n - \tilde{q}_n, \tag{4.1}$$

$$\tilde{q}_n = sign(\tilde{h} * v + y_n) \tag{4.2}$$

with the initial condition $v_n = 0$ *for* $n < 0$. *This modulator is stable for all input sequences* $y \in \mathcal{Y}_\mu$ *if*

$$\|h\|_{\ell^1} + (1-\mu)\|h - \tilde{h}\|_{\ell^1} \leq 2 - \mu. \tag{4.3}$$

Proof. We prove by induction that $|v_n| \leq V$ for all n where $V < \infty$ is some constant yet to be determined. The seed is given by the initial condition. For the induction step, recall the notation $\|w\|_{\ell^\infty}^{(n)} := \sup_{j<n} |w_j|$. In this notation, the induction hypothesis reads $\|v\|_{\ell^\infty}^{(n)} \leq V$. To bound v_n, we distinguish two cases.

If $\tilde{q}_n = q_n := sign\left((h * v)_n + y_n\right)$, then

$$v_n = (h * v)_n + y_n - sign\left((h * v)_n + y_n\right) \tag{4.4}$$

and

$$|v_n| \leq \max\left(|(h * v)_n + y_n| - 1, 1\right) \leq \max\left(\|h\|_{\ell^1}\|v\|^{(n)}, 1\right) + \mu - 1. \tag{4.5}$$

In order to conclude that $|v_n| \leq V$, we need that

$$\max\left(\|h\|_{\ell^1} V + \mu - 1, 1\right) \leq V, \tag{4.6}$$

which can be rewritten as

$$1 \leq V \leq \frac{1-\mu}{\|h\|_{\ell^1} - 1}. \tag{4.7}$$

Conversely, if $\tilde{q}_n \neq q_n$, then either $\tilde{q}_n = -1$ and

$$(\tilde{h} * v)_n + y_n \leq 0 \leq (h * v)_n + y_n, \tag{4.8}$$

or $\tilde{q}_n = 1$ and

$$(h * v)_n + y_n \leq 0 \leq (\tilde{h} * v)_n + y_n. \tag{4.9}$$

In the first case, we obtain

$$0 \leq (h * v)_n + y_n \leq [(h - \tilde{h}) * v]_n \tag{4.10}$$

Combining (4.10) with the equality $v_n = (h * v)_n + y_n - \tilde{q}_n = (h * v)_n + y_n + 1$, we obtain that $v_n \geq 0$ and

$$|v_n| = v_n \leq [(h - \tilde{h}) * v]_n + 1 \leq \|h - \tilde{h}\|_{\ell^1} \|v\|_{\ell^\infty}^{(n)} + 1. \tag{4.11}$$

In the second case (given by Inequality 4.9), the signs in the intermediate steps of this estimate are reversed, but eventually one obtains the same bound.

In both cases, in order to conclude that $|v_n| \leq V$, we need that

$$\|h - \tilde{h}\|_{\ell^1} V + 1 \leq V, \tag{4.12}$$

or equivalently

$$V \geq \frac{1}{1 - \|h - \tilde{h}\|_{\ell^1}}. \tag{4.13}$$

Note that the right hand side of Inequality (4.13) is always greater than 1. Thus a number V that simultaneously satisfies conditions (4.7) and (4.13) exists

if and only if

$$\frac{1}{1-\|h-\tilde{h}\|_{\ell^1}} \leq \frac{1-\mu}{\|h\|_{\ell^1}-1}, \tag{4.14}$$

which can be rewritten as

$$\|h\|_{\ell^1} + (1-\mu)\|h-\tilde{h}\|_{\ell^1} \leq 2-\mu. \tag{4.15}$$

On the other hand, if condition (4.15) holds, then for V that satisfies (4.7) and (4.13) we can conclude using condition (4.6) or condition (4.12) that $|v_n| \leq V$, which concludes the proof by induction. □

4.2 General formulation of the stability criterion

The stability criterion of Theorem 4.1 applies to $\Sigma\Delta$ modulators which do not employ the greedy quantization rule. In this section, we will show how after an appropriate change of variables, the criterion also allows us to make inferences about the stability of modulators which do employ the greedy quantization rule.

Theorem 4.2. *For $\bar{g} \in \ell^1$ that has a convolutional inverse $\bar{g}^{-1}$ and $\bar{h}$ such that $\delta^{(0)} - \bar{h} = \Delta^m \bar{g}$, the modulator given by*

$$v_n = (\bar{h} * v)_n + y_n - q_n, \tag{4.16}$$

$$q_n = sign\left((\bar{h} * v)_n + y_n\right). \tag{4.17}$$

is stable if there exists an auxiliary sequence $g \in \ell^1$ such that $h = \delta^{(0)} - \Delta^m g$ satisfies

$$\|h\|_{\ell^1} + (1-\mu)(\|\delta^{(0)} - g * \overline{g}^{-1}\|_{\ell^1}) \leq 2-\mu. \tag{4.18}$$

Proof. We will perform a change of variables $v \to \tilde{v}$ such that equations (4.16) and (4.17) take a form as in Theorem 4.1, i.e.,

$$\tilde{v}_n = (h * \tilde{v})_n + y_n - q_n, \tag{4.19}$$

$$q_n = sign\left((\tilde{h} * \tilde{v})_n + y_n\right) \tag{4.20}$$

for some $\tilde{h}$ yet to be determined. For Equation (4.19) to be equivalent to Equation (4.16), $\tilde{v}$ should be chosen such that

$$\left(\delta^{(0)} - h\right) * \tilde{v} = \left(\delta^{(0)} - \bar{h}\right) * v, \tag{4.21}$$

and so, as $h = \delta^{(0)} - \Delta^m g$ and $\bar{h} = \delta^{(0)} - \Delta^m \bar{g}$ with g, $\bar{g}$ both in ℓ^1,

$$\begin{aligned} \bar{g} * v &= g * \tilde{v}, \\ v &= \bar{g}^{-1} * g * \tilde{v}. \end{aligned} \tag{4.22}$$

Furthermore, for Equation (4.20) to be equivalent to Equation (4.17), $\tilde{h}$ should be such that

$$\tilde{h} * \tilde{v} = \bar{h} * v = \left(\delta^{(0)} - \Delta^m \bar{g}\right) * \bar{g}^{-1} * g * \tilde{v} = \left(\bar{g}^{-1} * g - \Delta^m g\right) * \tilde{v}. \tag{4.23}$$

We conclude that

$$\tilde{h} = \bar{g}^{-1} * g - \Delta^m g = \bar{g}^{-1} * g - \delta^{(0)} + h. \tag{4.24}$$

Thus by Theorem 4.1 the modulator given by (4.19) and (4.20) is stable – i.e., $\tilde{v}$

is bounded – if

$$\|h\|_{\ell^1} + (1-\mu)\|\delta^{(0)} - \overline{g}^{-1} * g\|_{\ell^1} \leq 2 - \mu. \tag{4.25}$$

Now, as $g, \bar{g} \in \ell^1$, the sequence $v = \bar{g}^{-1} * g * \tilde{v}$ is bounded if and only $\tilde{v}$ is, and Inequality (4.25) is also a condition for stability of the modulator given by Equations (4.16) and (4.17). This completes the proof of the theorem. □

4.3 Application to MCR ΣΔ modulators

Condition (4.18) in Theorem 4.2 implies that $\|h\|_{\ell^1} < 2 - \mu$. Hence the modulator with the auxiliary variable h as its filter coefficient vector and the greedy quantization rule will have better stability behavior than the modulator with coefficient vector $\bar{h}$. Consequently, Theorem 4.2 is expected to be particularly useful if additional constraints are imposed on the structure of $\bar{h}$. These constraints need not hold for the auxiliary vector h, and we can make inference from general stable modulators to those with the additional structure. Of particular interest are MCR ΣΔ modulators, where $\bar{g}^{-1} = a$ and consequently, Condition (4.18) takes a particularly simple form. In this case, Theorem 4.2 is directly reformulated as follows.

Theorem 4.3. *The MCR ΣΔ modulator given by*

$$v_n = (\bar{h} * v)_n + y_n - q_n, \tag{4.26}$$

$$q_n = sign((\bar{h} * v)_n + y_n) \tag{4.27}$$

with $\delta^{(0)} - \bar{h} = \Delta^m a^{-1}$ *for some* a *with polynomial generating function*

$$A = \prod_{j=1}^{m} (1 - \xi_j z) \tag{4.28}$$

for complex numbers ξ_j *with* $|\xi_j| < 1$ *is stable if there exists* $g \in \ell^1$ *such that* $h = \delta^{(0)} - \Delta^m g$ *satisfies*

$$\|h\|_{\ell^1} + (1 - \mu)\|\delta^{(0)} - a * g\|_{\ell^1} \leq 2 - \mu. \tag{4.29}$$

Remark: The basic stability criterion given in Theorem 1.6 is a special case Theorems 4.2 and 4.3; one can choose $g = \bar{g}$ or, when we work with an MCR modulator, $g = a^{-1}$. We do not know to which extent these results are true generalizations of the basic criterion, i.e., how many (if any) modulators exist the stability of which can be shown via Theorem 4.2 or Theorem 4.3, but not via the basic criterion.

Bibliography

[1] J. J. Benedetto, A. M. Powell, and Ö. Yılmaz. Sigma-Delta ($\Sigma\Delta$) quantization and finite frames. *IEEE Trans. Inform. Theory*, 52(5):1990–2005, 2006.

[2] P. Borwein and T. Erdélyi. *Polynomials and polynomial inequalities*, volume 161 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

[3] A. R. Calderbank and I. Daubechies. The pros and cons of democracy. *IEEE Trans. Inform. Theory*, 48(6):1721–1725, 2002. Special issue on Shannon theory: perspective, trends, and applications.

[4] I. Daubechies and R. DeVore. Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order. *Ann. of Math*, 158:679–710, 2003.

[5] R. A. DeVore and G. G. Lorentz. *Constructive Approximation*. Springer-Verlag, 1993.

[6] C. S. Güntürk. *Harmonic Analysis of Two Problems in Signal Quantization and Compression*. PhD thesis, Princeton University, 2000.

[7] C. S. Güntürk. One-bit sigma-delta quantization with exponential accuracy. *Comm. Pure Appl. Math.*, 56:1608–1630, 2003.

[8] C. S. Güntürk. Approximating a bandlimited function using very coarsely quantized data: improved error estimates in sigma-delta modulation. *J. Amer. Math. Soc.*, 17(1):229–242 (electronic), 2004.

[9] C. S. Güntürk and N. T. Thao. Ergodic dynamics in sigma-delta quantization: tiling invariant sets and spectral analysis of error. *Adv. in Appl. Math.*, 34(3):523–560, 2005.

[10] D. Haroske and H. Triebel. *Distributions, Sobolev Spaces, Elliptic Equations*. EMS, 2008.

[11] F. Krahmer. An improved family of exponentially accurate sigma-delta quantization schemes. In *Wavelets XII. Edited by Van De Ville, Dimitri; Goyal, Vivek K.; Papadakis, Manos. Proceedings of the SPIE*, volume 6701, October 2007.

[12] Y. Meyer. *Wavelets and Operators*. Cambridge University Press, 1992.

[13] S. R. Norsworthy, R. Schreier, and G. C. Temes, editors. *Delta-Sigma-Converters: Theory, Design and Simulation*. Wiley-IEEE, 1996.

[14] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-Time Signal Processing*. Prentice-Hall signal processing. Prentice-Hall, Upper Saddle River, NJ, second edition, 1999.

[15] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab. *Signals and systems*. Prentice Hall, Upper Saddle River, NJ, second edition, 1997.

[16] R. Schreier. An empirical study of high-order single-bit delta-sigma modulators. *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, 40(8):461–466, Aug 1993.

[17] R. Schreier and M. Snelgrove. Stability in a general sigma delta modulator. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pages 1769–1772, 1991.

[18] G. Szegö. *Orthogonal Polynomials*. Amer. Math. Soc., 1939.

[19] Ö Yılmaz. Stability analysis for several sigma-delta methods of coarse quantization of bandlimited functions. *Constructive Approximation*, 18(4):599–623, 2002.