An Optimal Family of Exponentially Accurate One-Bit Sigma-Delta Quantization Schemes

Percy Deift* C. Sinan Güntürk* Felix Krahmer*†

January 21, 2010

Abstract

Sigma-Delta modulation is a popular method for analog-to-digital conversion of bandlimited signals that employs coarse quantization coupled with oversampling. The standard mathematical model for the error analysis of the method measures the performance of a given scheme by the rate at which the associated reconstruction error decays as a function of the oversampling ratio λ. It was recently shown that exponential accuracy of the form $O(2^{-r\lambda})$ can be achieved by appropriate one-bit Sigma-Delta modulation schemes. By general information-entropy arguments r must be less than 1. The current best known value for r is approximately 0.088. The schemes that were designed to achieve this accuracy employ the "greedy" quantization rule coupled with feedback filters that fall into a class we call "minimally supported". In this paper, we study the minimization problem that corresponds to optimizing the error decay rate for this class of feedback filters. We solve a relaxed version of this problem exactly and provide explicit asymptotics of the solutions. From these relaxed solutions, we find asymptotically optimal solutions of the original problem, which improve the best known exponential error decay rate to $r \approx 0.102$. Our method draws from the theory of orthogonal polynomials; in particular, it relates the optimal filters to the zero sets of Chebyshev polynomials of the second kind.

1 Introduction

Conventional Analog-to-Digital (A/D) conversion systems consist of two basic steps: *sampling* and *quantization*. Sampling is the process of replacing the input function $(x(t))_{t\in\mathbb{R}}$ by a sequence of its sample values $(x(t_n))_{n\in\mathbb{Z}}$. The sampling instances (t_n) are typically uniform, i.e., $t_n = n\tau$ for some $\tau > 0$. It is well known that this process incurs no loss of information if the signal is bandlimited and τ is sufficiently small. More precisely, if supp $\widehat{x} \subset [-\Omega, \Omega]$, then it suffices to pick $\tau \leq \tau_{\text{crit}} := \frac{1}{2\Omega}$, and the *sampling theorem* provides a recipe for perfect

*Courant Institute of Mathematical Sciences, New York University, New York, NY, USA.
†Hausdorff Center for Mathematics, Universität Bonn, Bonn, Germany.

reconstruction via the formula

$$x(t) = \tau \sum_{n\in\mathbb{Z}} x(n\tau)\varphi(t - n\tau), \tag{1}$$

where φ is any sufficiently localized function such that

$$\widehat{\varphi}(\xi) = \begin{cases} 1, & |\xi| \le \Omega, \\ 0, & |\xi| \ge \frac{1}{2\tau}, \end{cases} \tag{2}$$

which we shall refer to as the *admissibility* condition for φ. Here, $\widehat{x}$ and $\widehat{\varphi}$ denote the Fourier transform of x and φ, respectively. In this paper, we shall work with the following normalization of the Fourier transform on $\mathbb{R}$:

$$\widehat{f}(\xi) := \int_{-\infty}^{\infty} f(t)e^{-2\pi i\xi t}\,\mathrm{d}t.$$

The value $\rho := 1/\tau$ is called the sampling rate, and $\rho_{\text{crit}} := 1/\tau_{\text{crit}} = 2\Omega$ is called the critical (or Nyquist) sampling rate. The *oversampling ratio* is defined as

$$\lambda := \frac{\rho}{\rho_{\text{crit}}}. \tag{3}$$

We shall assume in the rest of the paper that the value of Ω is arbitrary but fixed.

The next step, quantization, is the process of discretization of the amplitude, which involves replacing each $x(t_n)$ by another value q_n suitably chosen from a fixed, typically finite set $\mathcal{A}$ (the alphabet). The resulting sequence (q_n) forms the raw digital representation of the analog signal x.

The standard approach to recover an approximation $\tilde{x}$ to the original analog signal x is to use the reconstruction formula (1) with $x(n\tau)$ replaced by q_n, which produces an error signal $e := x - \tilde{x}$ given by

$$e(t) = \tau \sum_{n\in\mathbb{Z}} \big(x(n\tau) - q_n\big)\varphi(t - n\tau). \tag{4}$$

The quality of this approximation can then be measured by a variety of functional norms on the error signal. In this paper we will use the norm $\|e\|_{L^\infty}$ which is standard and arguably the most meaningful.

Traditionally, A/D converters have been classified into two main families: *Nyquist-rate* converters ($\lambda \approx 1$) and *oversampling* converters ($\lambda \gg 1$). When $\tau = \tau_{\text{crit}}$, the set of functions $\{\varphi(\cdot - n\tau) : n \in \mathbb{Z}\}$ forms an orthogonal system. Therefore, a Nyquist-rate converter necessarily has to keep $|x(n\tau) - q_n|$ small for each n in order to achieve small overall reconstruction error; this requires that the alphabet $\mathcal{A}$ forms a fine net in the range of the signal. On the other hand, oversampling converters are not bound by this requirement because as τ decreases (i.e., λ increases), the kernel of the operator

$$T_\tau^\varphi : (c_n) \mapsto \sum_{n\in\mathbb{Z}} c_n\varphi(\cdot - n\tau) \tag{5}$$

gets "bigger" in a certain sense, and small error can be achieved even with very coarse alphabets $\mathcal{A}$. The extreme case is a *one-bit quantization* scheme, which uses $\mathcal{A} = \{-1, +1\}$. For the circuit engineer, coarse alphabets mean 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 (see section 2.1 below) have become more popular than Nyquist-rate converters for low to medium-bandwidth signal classes, such as audio [12]. On the other hand, the mathematics of oversampled A/D conversion is highly challenging as the selection problem of (q_n) shifts from being local (i.e., the q_n's are chosen independently for each n) to a more global one; a quantized assignment $q_n \in \mathcal{A}$ should be computed based not only on the current sample value $x(t_n)$, but also taking into account sample values $x(t_k)$ and assignments q_k in neighboring positions. Many "online" quantization applications, such as A/D conversion of audio signals, require causality, i.e., only quantities that depend on prior instances of time can be utilized. Other applications, such as digital halftoning, may not be strictly bound by the same kind of causality restrictions although it is still useful to process samples in some preset order. In both situations, the amount of memory that can be employed in the quantization algorithm is one of the limiting factors determining the performance of the algorithm.

This paper is concerned with the approximation theory of oversampled, coarse quantization, in particular, one-bit quantization of bandlimited functions. Despite the vast engineering literature on the subject (e.g., see [12]), and a recent series of more mathematically oriented papers (e.g., [5, 15, 7, 8, 9, 1, 2]), the fundamental question of how to carry out optimal quantization remains open. After the pioneering work of Daubechies and DeVore on the mathematical analysis and design of $\Sigma\Delta$ modulators [5], more recent work showed that exponential accuracy in the oversampling ratio λ can be achieved by appropriate one-bit $\Sigma\Delta$ modulation schemes [7]. The best achievable error decay rate for these schemes was $O(2^{-r\lambda})$ with $r \approx 0.076$ in [7]. Later, with a modification, this rate was improved to $r \approx 0.088$ in [11]. It is known that any one-bit quantization scheme has to obey $r < 1$, whereas it is not known if this upper bound is tight [4, 7]. This paper improves the best achievable rate further to $r \approx 0.102$ by designing an optimal family of Sigma-Delta modulation schemes within the class of so-called "minimally supported" recursion filters. These schemes were introduced in [7] together with a number of accompanying open problems. The **main results** of this paper (see Theorems **4.1, 5.5, 5.6**) are based on the solution of one of these problems, namely, the optimization of the minimally supported recursion filters that are used in conjunction with the *greedy* rule of quantization. One of our main results is that the supports of these optimal filters are given by suitably scaled zero sets of Chebyshev polynomials of the second kind, and we use this result to derive our improved exponent for the error bound.

The paper is organized as follows. In Section 2, we review the basic mathematical theory of $\Sigma\Delta$ modulation as well as the constructions and methods of [7] which will be relevant to this paper, such as the family of minimally supported recursion filters, the greedy quantization rule, and fundamental aspects of the error analysis leading to exponentially decaying error bounds. As we explain, the discrete optimization problem introduced in [7] plays a key role in the analysis. In Section 3, we introduce a relaxed version of the optimization

problem, which is analytically tractable. This relaxed problem is solved in Section 4 and analyzed asymptotically in Section 5 as the order of the reconstruction filter goes to infinity. Section 5 also contains the construction of asymptotically optimal solutions to the discrete optimization problem, which yields the exponential error decay rate mentioned above. Finally, we extend our results to multi-level quantization alphabets in Section 6. We also collect, separately in the Appendix, the properties and identities for Chebyshev polynomials which are used in the proofs of our results.

2 Background on $\Sigma\Delta$ modulation

2.1 Noise shaping and feedback quantization

$\Sigma\Delta$ modulation is a generic name for a family of recursive quantization algorithms that utilize the concept of "noise shaping". (The origin of the terminology $\Sigma\Delta$, or alternatively $\Delta\Sigma$, goes back to the patent application of Inose et al [10] and refers to the presence of certain circuit components at the level of A/D circuit implementation; see also [12, 13].) Let (y_n) be a general sequence to be quantized (for example, $y_n = x(n\tau)$), (q_n) denote the quantized representation of this sequence, and (ν_n) be the quantization error sequence (i.e., the "noise"), defined by $\nu := y - q$. Noise shaping is a quantization strategy whose objective is to arrange for the quantization noise ν to fall outside the frequency band of interest, which, in our case is the low-frequency band. Note that the effective error we are interested in is $e = T_\tau^\varphi(\nu)$, hence one would like ν to be close to the kernel of T_τ^φ. It is useful to think of $T_\tau^\varphi(\nu)$ as a generalized convolution of the sequence ν and the function φ (sampled at the scale τ). Let us introduce the notation

$$\nu \circledast_\tau \varphi := T_\tau^\varphi(\nu). \tag{6}$$

Note that $(a * b) \circledast_\tau \varphi = a \circledast_\tau (b \circledast_\tau \varphi)$ and $a \circledast_\tau (\varphi * \psi) = (a \circledast_\tau \varphi) * \psi$, etc., where a and b are sequences on $\mathbb{Z}$, φ and ψ are functions on $\mathbb{R}$, and $*$ denotes the usual convolution operation (on $\mathbb{Z}$ and $\mathbb{R}$). By taking the Fourier transform of (5) one sees that the kernel of T_τ^φ consists of sequences ν that are spectrally disjoint from φ, i.e., "high-pass" sequences, since φ is a "low-pass" function, as apparent from (2). Thus, arranging for ν to be (close to) a high-pass sequence is the primary objective of $\Sigma\Delta$ modulation.

High-pass sequences have the property that their Fourier transforms vanish at zero frequency (and possibly in a neighborhood). For a finite high-pass sequence s, the Fourier transform $\widehat{s}(\xi) := \sum s_n e^{2\pi i n\xi}$ has a factor $(1 - e^{2\pi i\xi})^m$ for some positive integer m, which means that $s = \Delta^m w$ for some finite sequence w. Here, Δ denotes the finite difference operator defined by

$$(\Delta w)_n := w_n - w_{n-1}. \tag{7}$$

The quantization error sequence $\nu = y - q$, however, need not be finitely supported, and therefore $\widehat{\nu}$ need not be defined as a function (since ν is bounded at best). Nevertheless, this spectral factorization can be used to model more general high-pass sequences. Indeed, a $\Sigma\Delta$

modulation scheme of order m utilizes the difference equation

$$y - q = \Delta^m u \tag{8}$$

to be satisfied for each input y and its quantization q, for an appropriate auxiliary sequence u (called the state sequence). This explicit factorization of the quantization error is useful if u is a bounded sequence, as will be explained in more detail in the next subsection.

In practice, (8) is used as part of a quantization algorithm. That is, given any sequence $(y_n)_{n\geq 0}$, its quantization $(q_n)_{n\geq 0}$ is generated by a recursive algorithm that satisfies (8). This is achieved via an associated "quantization rule"

$$q_n = Q(u_{n-1}, u_{n-2}, \dots, y_n, y_{n-1}, \dots), \tag{9}$$

together with the "update rule"

$$u_n = \sum_{k=1}^{m} (-1)^{k-1} \binom{m}{k} u_{n-k} + y_n - q_n, \tag{10}$$

which is a restatement of (8). Typically one employs the initial conditions $u_n = 0$ for $n < 0$.

In the electrical engineering literature, such a recursive procedure for quantization is called "feedback quantization" due to the role q_n plays as a (nonlinear) feedback term via (9) for the difference equation (8). Note that if y and q are given unrelated sequences and u is determined from $y - q$ via (8), then u would typically be unbounded for $m \geq 1$. Hence the role of the quantization rule Q is to tie q to y in such a way as to control u.

A $\Sigma\Delta$ modulator may also arise from a more general difference equation of the form

$$y - q = H * v \tag{11}$$

where H is a *causal* sequence (i.e., $H_n = 0$ for $n < 0$) in ℓ^1 with $H_0 = 1$. If $H = \Delta^m g$ with $g \in \ell^1$, then any (bounded) solution v of (11) gives rise to a (bounded) solution u of (8) via $u = g * v$. Thus (11) can be rewritten in the *canonical* form (8) by a change of variables. Nevertheless, there are significant advantages to working directly with representations of the form (11), as explained in Section 2.3 below.

2.2 Basic error estimates

Basic error analysis of $\Sigma\Delta$ modulation only relies on the *boundedness* of solutions u of (8) for given y and q, and not on the specifics of the quantization rule Q that was employed. It is useful, however, to consider arbitrary quantizer maps $\mathcal{M} : y \mapsto q$ that satisfy (8) for some m and u, where u may or may not be bounded. Note that if u is a solution to (8) for a given triple (y, q, m), then $\tilde{u} := \Delta u$ is a solution for the triple $(y, q, m-1)$. Hence a given map $\mathcal{M}$ can be treated as a $\Sigma\Delta$ modulator of different orders. Formally, we will refer to the pair $(\mathcal{M}, m)$ as a $\Sigma\Delta$ modulator of order m.

As indicated above, in order for a $\Sigma\Delta$ modulator $(\mathcal{M}, m)$ to be useful, there must be a bounded solution u to (8). (Note that, up to an additive constant, there can be at most one

such bounded solution once $m > 0$.) Moreover, one would like this to be the case for all input sequences y in a given class $\mathcal{Y}$, such as $\mathcal{Y}_\mu = \{y : ||y||_{\ell^\infty} \leq \mu\}$ for some $\mu > 0$. In this case, we say that $(\mathcal{M}, m)$ is *stable* for the input class $\mathcal{Y}$. Clearly, if $(\mathcal{M}, m)$ is stable for $\mathcal{Y}$, then $(\mathcal{M}, m-1)$ is stable for $\mathcal{Y}$ as well. To any quantizer map $\mathcal{M}$ and a class of inputs $\mathcal{Y}$, we assign its maximal order $m^*(\mathcal{M}, \mathcal{Y})$ via

$$m^*(\mathcal{M}, \mathcal{Y}) := \sup\big\{m : \forall y \in \mathcal{Y}, \quad \exists u \in \ell^\infty \text{ such that } y - \mathcal{M}(y) = \Delta^m u\big\}. \tag{12}$$

Note that both 0 and ∞ are admissible values for $m^*(\mathcal{M}, \mathcal{Y})$. With this notation, $(\mathcal{M}, m)$ is stable for the class $\mathcal{Y}$ if and only if $m \leq m^*(\mathcal{M}, \mathcal{Y})$.

Stability is a crucial property. Indeed, it was shown in [5] that a stable m-th order scheme with bounded solution u results in the error bound

$$\|e\|_{L^\infty} \leq \|u\|_{\ell^\infty} \|\varphi^{(m)}\|_{L^1} \tau^m, \tag{13}$$

where $\varphi^{(m)}$ denotes the mth order derivative of φ. The proof of (13) employs repeated summation by parts, giving rise to the commutation relation

$$(\Delta^m u) \circledast_\tau \varphi = u \circledast_\tau (\Delta_\tau^m \varphi), \tag{14}$$

where Δ_τ is the finite difference operator at scale τ defined by $(\Delta_\tau \varphi)(\cdot) := \varphi(\cdot) - \varphi(\cdot - \tau)$. The left hand side of (14) is simply equal to the error signal e by definition. On the other hand, the right hand side of this relation immediately leads to the error bound (13). As u is not uniquely determined by the equation $y - \mathcal{M}(y) = \Delta^m u$, it is convenient to introduce the notation

$$U(\mathcal{M}, m, y) := \inf\big\{\|u\|_{\ell^\infty} : y - \mathcal{M}(y) = \Delta^m u\big\}, \tag{15}$$

and for an input class $\mathcal{Y}$

$$U(\mathcal{M}, m, \mathcal{Y}) := \sup_{y \in \mathcal{Y}} U(\mathcal{M}, m, y). \tag{16}$$

We are interested in applying (13) to sequences y that arise as samples $y_n = x(n\tau)$ of a bandlimited signal x as above. To compare the error bounds for different values of τ, one could consider the class $\mathcal{Y}^{(x)} = \{y = (y_n)_{n \in \mathbb{Z}} : y_n = x(n\tau) \text{ for some } \tau\}$ and work with the constant $U(\mathcal{M}, m, \mathcal{Y}^{(x)})$. However, it is difficult to estimate $U(\mathcal{M}, m, \mathcal{Y}^{(x)})$ in a way that accurately reflects the detailed nature of the signal x. Instead, we note that for $\|x\|_{L^\infty} \leq \mu$, one has $\mathcal{Y}^{(x)} \subset \mathcal{Y}_\mu$, where $\mathcal{Y}_\mu = \{y = (y_n)_{n \in \mathbb{Z}} : ||y||_{\ell^\infty} \leq \mu\}$ as defined above. This leads to the bound

$$\|e\|_{L^\infty} \leq U(\mathcal{M}, m, \mathcal{Y}_\mu) \|\varphi^{(m)}\|_{L^1} \tau^m. \tag{17}$$

In (17), the reconstruction kernel φ is restricted by the τ-dependent admissibility condition (2). However, if a reconstruction kernel φ_0 is admissible in the sense of (2) for $\tau = \tau_0 = 1/\rho_0$, then it is admissible for all $\tau < \tau_0$. This allows one to fix $\rho_0 = (1+\epsilon)\rho_{\text{crit}}$ for some small $\epsilon > 0$, and set $\varphi = \varphi_0$. Now Bernstein's inequality[1] implies that $\|\varphi_0^{(m)}\|_{L^1} \leq (\pi\rho_0)^m \|\varphi_0\|_{L^1}$

[1] If $\widehat{f}$ is supported in $[-A, A]$, then $\|f'\|_{L^p} \leq 2\pi A \|f\|_{L^p}$ for $1 \leq p \leq \infty$.

since $\widehat{\varphi}_0$ is supported in $[-\frac{\rho_0}{2}, \frac{\rho_0}{2}]$. This estimate allows us to express the error bound naturally as a function of the oversampling ratio λ defined in (3), as follows:

Let ϵ and φ_0 be fixed as above, let $\mathcal{M}$ be a given quantizer function, and let m be a positive integer. Then for any given Ω-bandlimited function x with $\|x\|_{L^\infty} \leq \mu$, the quantization error $e = e_\lambda$, as expressed in (4) in terms of $\tau = \frac{1}{\lambda \rho_{crit}}$, can be bounded in terms of λ:

$$\|e_\lambda\|_{L^\infty} \leq U(\mathcal{M}, m, \mathcal{Y}_\mu) \|\varphi_0\|_{L^1} \pi^m (1+\epsilon)^m \lambda^{-m}. \tag{18}$$

As the constants are independent of λ, this yields an $O(\lambda^{-m})$ bound as a function of λ. Note that any dependency of the error on the original bandwidth parameter Ω has now been effectively absorbed into the fixed constant $\|\varphi_0\|_{L^1}$. In fact, this quantity need not even depend on Ω; a reconstruction kernel φ_0^* can be designed once and for all corresponding to $\Omega = 1$, and then employed to define $\varphi_0(t) := \Omega \varphi_0^*(\Omega t)$, which is admissible and has the same L^1-norm as φ_0^*.

2.3 Exponentially accurate $\Sigma\Delta$ modulation

The $O(\lambda^{-m})$ error decay rate derived in the previous section is based on using a fixed $\Sigma\Delta$ modulator. It is possible, however, to improve the bounds by choosing the modulator adaptively as a function of λ. In order to obtain an error decay rate better than polynomial, one needs an infinite family $((\mathcal{M}_m, m))_1^\infty$ of stable $\Sigma\Delta$ modulation schemes from which the optimal scheme $\mathcal{M}_{m_{opt}}$ (i.e., one that yields the smallest bound in (18)) is selected as a function of λ. This point of view was first pursued systematically in [5]. Finding a stable $\Sigma\Delta$ modulator is in general a non-trivial matter as m increases, and especially so if the alphabet $\mathcal{A}$ is a small set. The extreme case is one-bit $\Sigma\Delta$ modulation, i.e., when $\mathrm{card}(\mathcal{A}) = 2$. The first infinite family of arbitrary-order, stable one-bit $\Sigma\Delta$ modulators was also constructed in [5]. In the one-bit case, one may set $\mathcal{A} = \{-1, +1\}$ to normalize the amplitude, and choose $\mu \leq 1$ when defining the input class $\mathcal{Y} = \mathcal{Y}_\mu$. The optimal order m_{opt} and the size of the resulting bound on $\|e_\lambda\|_{L^\infty}$ depend on the constants $U(\mathcal{M}_m, m, \mathcal{Y}_\mu)$.

There are a priori lower bounds on $U(\mathcal{M}, m, \mathcal{Y}_\mu)$ for any $0 < \mu \leq 1$ and any quantizer $\mathcal{M}$. Indeed, as shown in [7], one obtains a super-exponential lower bound on $U(\mathcal{M}, m, \mathcal{Y}_\mu)$ by considering the average metric entropy of the space of bandlimited functions, as follows. Define the quantity

$$U_m(\mathcal{Y}) := \inf_{\mathcal{M}} \; U(\mathcal{M}, m, \mathcal{Y}) \tag{19}$$

and let

$$\mathcal{X}_\mu := \{x : \mathrm{supp}\ \widehat{x} \in [-1/2, 1/2],\ \|x\|_{L^\infty} \leq \mu\}. \tag{20}$$

Then, as shown in [6, 7], one has for any one-bit quantizer

$$\sup\{\|e_\lambda\|_{L^\infty} : x \in \mathcal{X}_\mu\} \gtrsim_\mu 2^{-\lambda}, \tag{21}$$

which yields, when used with (18), the result that $U_m(\mathcal{Y}_\mu) \gtrsim_\mu (mc)^m$ for some absolute constant $c > 0$ [7]. Here by $A \gtrsim_s B$, we mean that $A \geq CB$ for some constant C that may depend on s, but not on any other input variables of A and B.

For the family of $\Sigma\Delta$ modulators constructed in [5], which we will denote by $(\mathcal{D}_m)_{m=1}^{\infty}$, the best upper bounds on $U(\mathcal{D}_m, m, \mathcal{Y}_\mu)$ are of order $\exp(cm^2)$, resulting in the order-optimized error bound $\|e_\lambda\|_{L^\infty} \lesssim \exp(-c(\log\lambda)^2)$, which is substantially larger than the exponentially small lower bound of (21). The first construction that led to exponentially accurate $\Sigma\Delta$ modulation was given later in [7]. The associated modulators, here denoted by $(\mathcal{G}_m)_{m=1}^{\infty}$, satisfy the bound

$$U(\mathcal{G}_m, m, \mathcal{Y}_\mu) \lesssim (ma)^m, \tag{22}$$

for $a = a(\mu) > 0$. Substituting (22) into (18) and using the elementary inequality

$$\min_m \; m^m \alpha^{-m} \lesssim e^{-\alpha/e} \tag{23}$$

with $\alpha = \lambda/(a\pi(1+\epsilon))$, one obtains the order-optimized error bound

$$\sup\{\|e_\lambda\|_{L^\infty} : x \in \mathcal{X}_\mu\} \lesssim 2^{-r\lambda}, \tag{24}$$

where $r = r(\mu) = (a\pi e(1+\epsilon)\log 2)^{-1}$. On the other hand, the smallest achievable value of a is shown to be $6/e$, which corresponds to the largest achievable value of $r = r_{\max}\left((\mathcal{G}_m)_{m=1}^{\infty}\right) \approx (6\pi\log 2)^{-1} \approx 0.076$. Observe from (21) that it is impossible to achieve an exponent $r > 1$.

In [7], the $\mathcal{G}_m$ were given in the form (11) where the causal sequences $H, g \in \ell^1$ with $H_0 = g_0 = 1$ depend on m, and are related via

$$H = \Delta^m g. \tag{25}$$

Define $h := \delta^{(0)} - H$, where $\delta^{(0)}$ denotes the Kronecker delta sequence supported at 0. Then (11) can be implemented recursively as

$$v_n = (h * v)_n + y_n - q_n. \tag{26}$$

If q is chosen so that the resulting v is bounded, i.e., if this new scheme is stable, then $u := g * v$ is a bounded solution of (8), and

$$\|u\|_{\ell^\infty} \le \|g\|_{\ell^1}\|v\|_{\ell^\infty}. \tag{27}$$

The significance of the more general form (11) giving rise to (26) is the following: If

$$||h||_1 + ||y||_\infty \le 2, \tag{28}$$

then the *greedy* quantization rule

$$q_n := \operatorname{sign}\left((h * v)_n + y_n\right). \tag{29}$$

leads to a solution v with $||v||_\infty \le 1$, provided the initial conditions satisfy this bound. However, for the canonical form (8), the condition (28) clearly fails for all $m > 1$.

For $\|y\|_\infty \le \mu \le 1$, a filter h will satisfy (28) and hence lead to a stable scheme when $\|h\|_1 \le 2 - \mu$. Note that (25) together with the fact that $g \in \ell^1$ implies that $\sum h_i = 1$ and hence $\|h\|_1 \ge 1$.

In view of the preceding considerations, we are led, for each m to the following minimization problem:

$$\text{Minimize } \|g\|_1 \text{ subject to } \delta^{(0)} - h = \Delta^m g,\ \|h\|_1 \leq 2 - \mu. \tag{30}$$

It was shown in [7] that if a sequence of filters $h^{(m)}$ satisfy the *feasibility conditions* in (30) for the corresponding $m \in \mathbb{N}$, the diameter of the support set of $h^{(m)}$ must grow at least quadratically in m.

The minimization problem (30) was not solved in [7]; rather the author introduced a class of feasible filters $h = h^{(m)}$ which were effective in the sense that they lead to an exponential error bound with rate constant r as above. 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 can be shown to be the minimal support size for which $h^{(m)}$ can satisfy the feasibility conditions. We shall call such filters *minimally supported*. Note that if h has finite support and $\|g\|_1 < \infty$, then g has finite support. We make the following formal definition:

Definition 2.1. *We say that a filter $h = \delta^{(0)} - \Delta^m g$, for a finitely supported g, has minimal support if $|supp\ h| = m$.*

The goal of this paper is to find optimal filters within the class of filters with minimal support.

2.4 Filters with minimal support

As the filter $\delta^{(0)} - h$ arises as the m-th order finite difference of the vector g, its entries have to satisfy m moment conditions. This implies that the support size of h is at least m, as advertised above.

For filters h with minimal support

$$h = \sum_{j=1}^{m} d_j \delta^{(n_j)}, \tag{31}$$

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 strict causality of h.

Indeed, one finds

$$d_j = \prod_{i=1}^{m}{}' \frac{n_i}{n_i - n_j}. \tag{32}$$

Here the notation $\prod'$, and analogously $\sum'$, indicates 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 $\|h\|_1 \leq 2 - \mu$ then takes the form

$$\sum_{j=1}^{m} \prod_{i=1}^{m}{}' \frac{n_i}{|n_i - n_j|} \leq 2 - \mu. \tag{33}$$

Furthermore, explicit computations lead to the identity

$$\|g\|_1 = \frac{\prod_{j=1}^m n_j}{m!}. \tag{34}$$

In this notation, minimization problem (30) takes the form

$$\text{Minimize } \frac{\prod_{j=1}^m n_j}{m!} \text{ over } \{\mathbf{n} = (n_1, \ldots, n_m) \in \mathbb{N}^m : (33) \text{ holds and } 1 \le n_1 < \cdots < n_m\} \tag{35}$$

For $\mu = 1$, Problem (35) 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, $j = 1, \ldots, m$, that satisfy (33). In particular, for $n_j(\sigma) = 1 + \sigma(j-1)$, one easily sees that

$$\lim_{\sigma\to\infty} \sum_{j=1}^m \prod_{i=1}^{m\,\prime} \frac{n_i(\sigma)}{|n_i(\sigma) - n_j(\sigma)|} \quad = \quad 1. \tag{36}$$

So for every $\mu < 1$, $\mathbf{n}(\sigma)$ satisfies constraint (33) for all σ large enough.

Furthermore, any minimizer $\mathbf{n}$ of problem (35) 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 $j = 1, \ldots, m$. Calculate

$$\sum_{j=1}^m \prod_{i=1}^{m\,\prime} \frac{\tilde{n}_i}{|\tilde{n}_i - \tilde{n}_j|} = \sum_{j=1}^m \prod_{i=1}^{m\,\prime} \frac{n_i - 1}{|n_i - n_j|} < \sum_{j=1}^m \prod_{i=1}^{m\,\prime} \frac{n_i}{|n_i - n_j|} \le 2 - \mu \tag{37}$$

and

$$\frac{\prod_{j=1}^m \tilde{n}_j}{m!} < \frac{\prod_{j=1}^m n_j}{m!}. \tag{38}$$

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

Hence, we can fix $n_1 \equiv 1$, which reduces problem (35) to minimizing

$$\eta(\mathbf{n}) := \prod_{j=2}^m n_j \tag{39}$$

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

$$\sum_{j=1}^m \prod_{i=1}^{m\,\prime} \frac{n_i}{|n_i - n_j|} \le \gamma, \tag{40}$$

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 $\gamma \le 2$.

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

3 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

$$\eta(\mathbf{x}) := \prod_{j=1}^{m-1} x_j \tag{41}$$

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

$$f(\mathbf{x}) := \sum_{j=0}^{m-1} \prod_{i=0}^{m-1}{}' \frac{x_i}{|x_i - x_j|} \quad \leq \quad \gamma, \tag{42}$$

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

Proof. A simple calculation shows that

$$F(\mathbf{r}) = \sum_{j=0}^{m-1} \prod_{i=0}^{m-1}{}' \frac{x_i}{|x_i - x_j|} = \sum_{j=0}^{m-1} \prod_{i<j} \frac{1}{r_{i+1} r_{i+2} \cdots r_j - 1} \prod_{i>j} \frac{1}{1 - \frac{1}{r_{j+1} r_{j+2} \cdots r_i}}, \tag{43}$$

from which the monotonicity is immediate. □

Definition 3.2. *If $\mathbf{x}, \mathbf{y} \in D$ and $1 \leq \frac{y_1}{x_1} \leq \cdots \leq \frac{y_{m-1}}{x_{m-1}}$, 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 3.1 is equivalent to the following:

Corollary 3.3. *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 (41), (42), 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 $\tilde{x}_j(t) = (1-t)x_j + t x_0$. Since $f \circ \tilde{\mathbf{x}}$ is continuous in t and

$$f(\tilde{\mathbf{x}}(0)) = f(\mathbf{x}) < \gamma, \tag{44}$$

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

$$\eta(\tilde{\mathbf{x}}(t)) = \prod_{j=0}^{m-1} ((1-t)x_j + tx_0)) \tag{45}$$

is decreasing in t, so

$$\eta(\tilde{\mathbf{x}}(t)) < \eta(\mathbf{x}), \tag{46}$$

and $\mathbf{x}$ cannot be a minimizer. Hence we can replace constraint (42) by the equality

$$f(\mathbf{x}) = \sum_{j=0}^{m-1} \prod_{i=0}^{m-1}{}' \frac{x_i}{|x_i - x_j|} = \gamma. \tag{47}$$

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

$$\frac{\partial}{\partial x_k} \frac{x_k}{|x_k - x_j|} = -x_j \frac{1}{x_k - x_j} \frac{1}{|x_k - x_j|}. \tag{48}$$

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

$$\begin{aligned} \frac{\partial}{\partial x_k} \prod_{i=0}^{m-1}{}' \frac{x_i}{|x_i - x_j|} &= \left(\prod_{\substack{i=0 \\ i\neq k}}^{m-1}{}' \frac{x_i}{|x_i - x_j|} \right) \left(-x_j \frac{1}{x_k - x_j} \frac{1}{|x_k - x_j|} \right) & (49) \\ &= -\frac{\eta(\mathbf{x})}{x_k} \frac{(-1)^j}{x_k - x_j} b_j, & (50) \end{aligned}$$

where we set from now on

$$b_j(\mathbf{x}) = \prod_{i=0}^{m-1}{}' \frac{1}{x_i - x_j}. \tag{51}$$

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

Furthermore, for $j = k$,

$$\frac{\partial}{\partial x_k} \prod_{i=0}^{m-1}{}' \frac{x_i}{|x_i - x_k|} = -\sum_{l=0}^{m-1}{}' \frac{\eta(\mathbf{x})}{x_k} \frac{(-1)^k}{x_k - x_l} b_k(\mathbf{x}). \tag{52}$$

Hence

$$\frac{\partial f}{\partial x_k} = -\frac{1}{x_k} \eta(\mathbf{x}) \sum_{j=0}^{m-1}{}' \frac{1}{x_k - x_j} \left((-1)^k b_k(\mathbf{x}) + (-1)^j b_j(\mathbf{x}) \right). \tag{53}$$

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

$$\frac{\partial f}{\partial x_{m-1}} \;<\; 0 \tag{54}$$

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

We now show that the infimum of η subject to (47) 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,

$$\|\mathbf{x}^{(n)}\|_\infty = |\mathbf{x}_{m-1}^{(n)}| \le \eta\left(\mathbf{x}^{(n)}\right) \le M, \tag{55}$$

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

$$\gamma = f(\mathbf{x}^{(n)}) \ge \prod_{i=0}^{m-1}{}' \frac{x_i^{(n)}}{|x_i^{(n)} - x_j^{(n)}|} \ge \frac{1}{M^{m-2}|x_j^{(n)} - x_k^{(n)}|}, \tag{56}$$

which implies that $|x_j^{(n)} - x_k^{(n)}| \ge \frac{1}{\gamma M^{m-2}} > 0$. It follows that $\mathbf{x}^{(n)}$ stays away from the boundary of D, which implies that $\mathbf{x}^{(\infty)} \in D$. Thus, problem (41), (47) 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 (41) and (47) solves the associated Lagrange multiplier equations, i.e., there exists $\nu = \nu(\mathbf{x}_{min}) \in \mathbb{R}$ such that

$$\nu \nabla\eta(\mathbf{x}_{min}) + \nabla f(\mathbf{x}_{min}) \;=\; 0, \tag{57}$$

$$f(\mathbf{x}_{min}) \;=\; \gamma. \tag{58}$$

Combined with (53) and the relation $\frac{\partial}{\partial y_k}\eta(\mathbf{y}) = \frac{1}{y_k}\eta(\mathbf{y})$, the Lagrange multiplier equations (57), (58) take the explicit form

$$\sum_{j=0}^{m-1}{}' \frac{1}{x_k - x_j}\left((-1)^k b_k(\mathbf{x}_{min}) + (-1)^j b_j(\mathbf{x}_{min})\right) \;=\; \nu, \tag{59}$$

$$f(\mathbf{x}_{min}) \;=\; \gamma \tag{60}$$

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 (59), (60) for some ν. In the following section, we will show that in fact η has a unique critical point in D.

4 Solution of the relaxed minimization problem

Theorem 4.1. *The minimum value of η on the manifold $\{f = \gamma\}$ in D is given by*

$$\eta = \eta_{min} = \frac{\sinh(2m\beta)}{(2\sinh\beta)^{2m-1}\cosh\beta} \tag{61}$$

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

$$\frac{\cosh\left((2m-1)\beta\right)}{\cosh\beta} = \gamma. \tag{62}$$

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

$$x_j = 1 + \frac{1}{2\sinh^2\beta}\left(1 + z_j\right), \quad j = 1, \dots, m-1. \tag{63}$$

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

Proof. The minimization problem (41), (47) assumes its minimum in D, so there must be at least one critical point $\mathbf{x}_{crit} = (x_1, \dots, x_{m-1})$ 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 4.2, that the equation can have only the solution given by (63).

As in (59), $\mathbf{x}_{crit} = (x_1, \dots, x_{m-1})$ must satisfy

$$\sum_{j=0}^{m-1}{}' \frac{1}{x_k - x_j}\left((-1)^k b_k(\mathbf{x}_{crit}) + (-1)^j b_j(\mathbf{x}_{crit})\right) = \nu(\mathbf{x}_{crit}), \tag{64}$$

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

In matrix notation, the statement reads

$$B(\mathbf{x}_{crit})\mathbf{v} = \nu(\mathbf{x}_{crit})\mathbf{e}, \tag{65}$$

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

$$B(\mathbf{y}) = \begin{pmatrix} \frac{b_0(\mathbf{y})}{y_1 - y_0} & \sum_{j=0}^{m-1}{}' \frac{b_1(\mathbf{y})}{y_1 - y_j} & \frac{b_2(\mathbf{y})}{y_1 - y_2} & \dots & \frac{b_{m-1}(\mathbf{y})}{y_1 - y_{m-1}} \\ \frac{b_0(\mathbf{y})}{y_2 - y_0} & \frac{b_1(\mathbf{y})}{y_2 - y_1} & \sum_{j=0}^{m-1}{}' \frac{b_2(\mathbf{y})}{y_2 - y_j} & \dots & \frac{b_{m-1}(\mathbf{y})}{y_2 - y_{m-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{b_0(\mathbf{y})}{y_{m-1} - y_0} & \frac{b_1(\mathbf{y})}{y_{m-1} - y_1} & \frac{b_2(\mathbf{y})}{y_{m-1} - y_2} & \dots & \sum_{j=0}^{m-1}{}' \frac{b_{m-1}(\mathbf{y})}{y_{m-1} - y_j} \end{pmatrix}, \tag{66}$$

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

$$p'_{\mathbf{y}}(s) = \prod_{j=1}^{m-1} (s - y_j). \tag{67}$$

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

$$I_k = \frac{1}{2\pi i} \oint_{\Gamma} \frac{p_{\mathbf{y}}(z)}{(z - y_k)(z - y_0)p'_{\mathbf{y}}(z)} dz, \quad k = 1, \ldots, m-1 \tag{68}$$

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

$$R_j = (-1)^{m-1} \frac{b_j(\mathbf{y})}{y_j - y_k} p(y_j). \tag{69}$$

At z_k, we have a double root in the denominator of the integrand in (68), so

$$\begin{aligned} R_k = \left(\frac{p_{\mathbf{y}}(z)}{\prod\limits_{\substack{i=0 \\ i \neq k}}^{m-1} (z - y_i)} \right)' \Bigg|_{z=y_k} &= \frac{p'_{\mathbf{y}}(y_k)}{\prod\limits_{\substack{i=0 \\ i \neq k}}^{m-1} (y_k - y_i)} - p_{\mathbf{y}}(y_k) \sum_{\substack{j=0 \\ j \neq k}}^{m-1} \frac{\prod\limits_{\substack{i=0 \\ i \neq j,k}}^{m-1} (y_k - y_i)}{\left(\prod\limits_{\substack{i=0 \\ i \neq k}}^{m-1} (y_k - y_i) \right)^2} \qquad (70) \\ &= (-1)^{m-1} {\sum_{j=0}^{m-1}}' \frac{b_k(\mathbf{y})}{y_j - y_k} p_{\mathbf{y}}(y_k) \qquad (71) \end{aligned}$$

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

$$\frac{(-1)^{m-1}}{m} = {\sum_{j=0}^{m-1}}' \frac{1}{y_j - y_k} \left(b_k(\mathbf{y}) p_{\mathbf{y}}(y_k) + b_j(\mathbf{y}) p_{\mathbf{y}}(y_j) \right) \tag{72}$$

or equivalently

$$B(\mathbf{y}) \mathbf{p}_{\mathbf{y}} = \frac{(-1)^m}{m} \mathbf{e}, \tag{73}$$

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 (73) 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 4.2, 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

$$B(\mathbf{x}_{crit})\mathbf{p}_{\mathbf{x}_{crit}} = \frac{(-1)^m}{m}\mathbf{e}, \tag{74}$$

and it follows that

$$B(\mathbf{x}_{crit})\left[m(-1)^m\nu(\mathbf{x}_{crit})\mathbf{p}_{\mathbf{x}_{crit}}\right] = \nu(\mathbf{x}_{crit})\mathbf{e}. \tag{75}$$

By (65), $\mathbf{v}$ also solves (75), and thus

$$\mathbf{v} - m(-1)^m\nu(\mathbf{x}_{crit})\mathbf{p}_{\mathbf{x}_{crit}} = c\,\mathbf{e} \tag{76}$$

for some constant c.

Set

$$q(s) = m(-1)^m\nu(\mathbf{x}_{crit})p_{\mathbf{x}_{crit}}(s) + c. \tag{77}$$

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

$$u(s) = (-1)^m q\left(\frac{x_m - 1}{2}(s+1) + 1\right) \tag{78}$$

has the equi-oscillation property, and we conclude by Proposition A.1 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

$$x_j = \frac{x_m - 1}{2}(1 + z_j) + 1. \tag{79}$$

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

$$x_j = x_j(K) := 1 + K(1 + z_j) \tag{80}$$

for some constant K. It follows from Lemma 3.1 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 (63).

We now compute $K = K(m, \gamma)$. The calculation uses several facts about Chebyshev polynomials and their roots, which we have collected in the Appendix. Let $\beta = \beta(m, \gamma) > 0$ be defined through the relation

$$K = \frac{1}{2\sinh^2\beta}. \tag{81}$$

Noting that $z_i = -z_{m-i}$, we obtain from Lemma A.2

$$f(\mathbf{x}) = \prod_{i=1}^{m-1} \frac{1+K(1+z_i)}{K\,|z_i - z_0|} + \sum_{j=1}^{m-1} \prod_{\substack{i=0 \\ i\neq j}}^{m-1} \frac{1+K(1+z_i)}{K\,|z_i - z_j|} \tag{82}$$

$$= \frac{2^{m-1}}{m} \prod_{i=1}^{m-1} \frac{1+K(1+z_i)}{K} + \sum_{j=1}^{m-1} \frac{2^{m-1}(1-z_j)}{m} \prod_{\substack{i=0 \\ i\neq j}}^{m-1} \frac{1+K(1+z_i)}{K} \tag{83}$$

$$= \left[\frac{2^{m-1}}{m} \prod_{i=1}^{m-1} \left(\frac{1}{K} + 1 - z_{m-i}\right)\right] \left[1 + \sum_{j=1}^{m-1} \frac{1+z_{m-j}}{1+K(1-z_{m-j})}\right] \tag{84}$$

$$= \left[\frac{2^{m-1}}{m} \prod_{i=1}^{m-1} \left(\frac{1}{K} + 1 - z_{i}\right)\right] \left[1 + \frac{1}{K} \sum_{j=1}^{m-1} \frac{1+z_{j}}{\left(1+\frac{1}{K}\right)-z_j}\right]. \tag{85}$$

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

$$\frac{2^{m-1}}{m} \prod_{i=1}^{m-1} \left(\frac{1}{K} + 1 - z_i\right) = \frac{1}{m^2} T'_m\left(1+\frac{1}{K}\right) = \frac{1}{m^2} T'_m\left(\cosh(2\beta)\right) = \frac{\sinh(2m\beta)}{m\sinh(2\beta)}. \tag{86}$$

Furthermore, differentiating (162), we obtain for $z = \cosh(\tau)$

$$\sum_{j=1}^{m-1} \frac{1}{z - z_j} = \frac{T''_m(z)}{T'_m(z)} = \frac{m\coth(m\tau) - \coth(\tau)}{\sinh\tau}. \tag{87}$$

Hence

$$1 + \frac{1}{K} \sum_{j=1}^{m-1} \frac{1+z_j}{\left(1+\frac{1}{K}\right)-z_j} \tag{88}$$

$$= 1 + (\cosh(2\beta)-1) \sum_{j=1}^{m-1} \left(-1 + \frac{1+\cosh(2\beta)}{\cosh(2\beta)-z_j}\right) \tag{89}$$

$$= 1 - (m-1)(\cosh(2\beta)-1) + (\cosh^2(2\beta)-1) \sum_{j=1}^{m-1} \frac{1}{\cosh(2\beta)-z_j} \tag{90}$$

$$= 1 - (m-1)(\cosh(2\beta)-1) + \sinh^2(2\beta) \frac{m\coth(2m\beta) - \coth(2\beta)}{\sinh 2\beta} \tag{91}$$

$$= m\left(1 - \cosh(2\beta) + \sinh(2\beta)\coth(2m\beta)\right). \tag{92}$$

Combining (92) and (86) yields

$$\begin{aligned}\gamma = f(\mathbf{x}) &= \frac{\sinh(2m\beta) - \cosh(2\beta)\sinh(2m\beta) + \sinh(2\beta)\cosh(2m\beta)}{\sinh(2\beta)} && (93)\\ &= \frac{\sinh(2m\beta) - \sinh((2m-2)\beta)}{2\sinh(\beta)\cosh(\beta)} && (94)\\ &= \frac{2\cosh((2m-1)\beta)\sinh(\beta)}{2\sinh(\beta)\cosh(\beta)} && (95)\\ &= \frac{\cosh((2m-1)\beta)}{\cosh(\beta)}, && (96)\end{aligned}$$

which proves (62). 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 (86), we write

$$\eta_{min} = \prod_{i=0}^{m-1}(1 + K(1+z_i)) = K^{m-1}\prod_{i=0}^{m-1}\left(\frac{1}{K} + 1 + z_i\right) = \frac{\sinh(2m\beta)}{(2\sinh(\beta))^{2m-1}\cosh(\beta)} \qquad (97)$$

□

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 4.2 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.

$$\tilde{B}(\mathbf{y}) = \begin{pmatrix} \sum_{l=0}^{m-1}{}' \frac{1}{y_0 - y_l} & \frac{1}{y_0 - y_1} & \frac{1}{y_0 - y_2} & \cdots & \frac{1}{y_0 - y_{m-1}} \\ \frac{1}{y_1 - y_0} & \sum_{l=0}^{m-1}{}' \frac{1}{y_1 - y_l} & \frac{1}{y_1 - y_2} & \cdots & \frac{1}{y_1 - y_{m-1}} \\ \frac{1}{y_2 - y_0} & \frac{1}{y_2 - y_1} & \sum_{l=0}^{m-1}{}' \frac{1}{y_2 - y_l} & \cdots & \frac{1}{y_2 - y_{m-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{y_{m-1} - y_0} & \frac{1}{y_{m-1} - y_1} & \frac{1}{y_{m-1} - y_2} & \cdots & \sum_{l=0}^{m-1}{}' \frac{1}{y_{m-1} - y_l} \end{pmatrix} \qquad (98)$$

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

$$J = \begin{pmatrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{99}$$

Proposition 4.2. *For $m \geq 2$, $\tilde{B}(\mathbf{y})$ has the Jordan decomposition*

$$\tilde{B}(\mathbf{y}) = P(\mathbf{y}) J P(\mathbf{y})^{-1}, \tag{100}$$

where

$$P(\mathbf{y}) = \begin{pmatrix} b_0(\mathbf{y}) & b_0(\mathbf{y})(y_0 - y_{m-1}) & \cdots & b_0(\mathbf{y})\frac{(y_0-y_{m-1})^{m-1}}{(m-1)!} \\ b_1(\mathbf{y}) & b_1(\mathbf{y})(y_1 - y_{m-1}) & \cdots & b_1(\mathbf{y})\frac{(y_1-y_{m-1})^{m-1}}{(m-1)!} \\ b_2(\mathbf{y}) & b_2(\mathbf{y})(y_2 - y_{m-1}) & \cdots & b_2(\mathbf{y})\frac{(y_2-y_{m-1})^{m-1}}{(m-1)!} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m-2}(\mathbf{y}) & b_{m-2}(\mathbf{y})(y_{m-2} - y_{m-1}) & \cdots & b_{m-2}(\mathbf{y})\frac{(y_{m-2}-y_{m-1})^{m-1}}{(m-1)!} \\ b_{m-1}(\mathbf{y}) & 0 & \cdots & 0 \end{pmatrix}. \tag{101}$$

Here the $b_j(\mathbf{y})$'s are defined as in (51).

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 (100) is equivalent to showing that $\tilde{B}(\mathbf{y})P(\mathbf{y}) = P(\mathbf{y})J$, that is, for $0 \leq j, n \leq m-1$,

$$\sum_{\substack{k=0 \\ k\neq j}}^{m-1} \frac{b_k(\mathbf{y})}{y_j - y_k} \frac{(y_k - y_{m-1})^n}{n!} + \sum_{\substack{l=0 \\ l\neq j}}^{m-1} \frac{b_j(\mathbf{y})}{y_j - y_l} \frac{(y_j - y_{m-1})^n}{n!} = b_j(\mathbf{y}) \frac{(y_j - y_{m-1})^{n-1}}{(n-1)!}, \tag{102}$$

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

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

$$J_{m,n}(t) = \frac{1}{2\pi i} \oint_{\Gamma} \frac{(z-t)^n}{\prod_{i=0}^{m-1}(z - y_i)} dz, \qquad 0 \leq n \leq m-1 \tag{103}$$

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

$$\delta^{(0)}_{n-(m-1)} = J_{m,n} = (-1)^{m-1} \sum_{k=0}^{m-1} b_k(\mathbf{y})(y_k - t)^n. \tag{104}$$

Now

$$\frac{\partial b_k}{\partial y_j} = \begin{cases} \frac{b_k(\mathbf{y})}{y_k - y_j} & \text{for } j \neq k \\ \sum\limits_{\substack{l=0 \\ l\neq j}}^{m-1} \frac{b_j(\mathbf{y})}{y_l - y_j} & \text{for } j = k \end{cases} \tag{105}$$

Hence, differentiating (104) with respect to y_j, leads to the identity

$$\sum_{\substack{k=0 \\ k\neq j}}^{m-1} \frac{b_k(\mathbf{y})}{y_k - y_j}(y_k - t)^n + \sum_{\substack{l=0 \\ l\neq j}}^{m-1} \frac{b_j(\mathbf{y})}{y_l - y_j}(y_l - t)^n + b_l(\mathbf{y}) n (y_j - t)^{n-1} = 0. \tag{106}$$

Letting $t \to y_{m-1}$, one obtains (102). □

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 5.1. (a) *For $K = K(m, \gamma)$ as in (80)*

$$\frac{2(m-1)^2}{(\cosh^{-1}\gamma)^2} - 1 \leq K \leq \frac{2m^2}{(\cosh^{-1}\gamma)^2}. \tag{107}$$

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

$$x_j^{(m)} \leq 1 + \sigma j^2. \tag{108}$$

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

$$\lim_{m\to\infty} x_j^{(m)} = 1 + \sigma j^2. \tag{109}$$

(d)

$$\lim_{m\to\infty} \frac{\left(\eta(\mathbf{x}^{(m)})\right)^{1/m}}{m^2} = \frac{1}{(\cosh^{-1}\gamma)^2} = \frac{\sigma}{\pi^2}. \tag{110}$$

Proof. We first provide bounds on β defined in (62). For a lower bound, write

$$\gamma = \frac{\cosh(2m-1)\beta)}{\cosh\beta} = \cosh(2m\beta) - \sinh(2m\beta)\tanh\beta \leq \cosh(2m\beta). \tag{111}$$

For an upper bound, we have

$$\gamma = \frac{\cosh(2m-1)\beta)}{\cosh\beta} = \cosh((2m-2)\beta) + \sinh((2m-2)\beta)\tanh\beta \geq \cosh((2m-2)\beta). \quad (112)$$

We obtain the bounds

$$\frac{1}{2m}\cosh^{-1}\gamma \leq \beta \leq \frac{1}{2m-2}\cosh^{-1}\gamma. \quad (113)$$

This implies the upper bound for K

$$K = \frac{1}{2\sinh^2\beta} \leq \frac{1}{2\beta^2} \leq \frac{2m^2}{(\cosh^{-1}\gamma)^2}. \quad (114)$$

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

$$K = \frac{1}{2\sinh^2\beta} \geq \frac{1}{2\beta^2} - 1 \geq \frac{2(m-1)^2}{(\cosh^{-1}\gamma)^2} - 1. \quad (115)$$

This proves (a).

Also

$$x_j^{(m)} = 1 + 2K\sin^2\left(\frac{j\pi}{2m}\right) \leq 1 + \frac{4m^2}{(\cosh^{-1}\gamma)^2}\left(\frac{j\pi}{2m}\right)^2 = 1 + \sigma j^2, \quad (116)$$

which proves (b).

From (107)

$$\lim_{m\to\infty}\frac{K(m,\gamma)}{m^2} = \frac{2}{(\cosh^{-1}\gamma)^2}, \quad (117)$$

and so

$$\lim_{m\to\infty} x_j^{(m)} = 1 + \frac{2}{(\cosh^{-1}\gamma)^2}\lim_{m\to\infty} 2m^2\sin^2\left(\frac{j\pi}{2m}\right) = 1 + \sigma j^2, \quad (118)$$

which proves (c).

Finally, from (113) we see that

$$\lim_{m\to\infty} 2m\beta(m,\gamma) = \cosh^{-1}\gamma, \quad (119)$$

and hence from (61)

$$\lim_{m\to\infty}\frac{(\eta(\mathbf{x}))^{1/m}}{m^2} = \lim_{m\to\infty}\frac{1}{m^2}\left(\frac{\sinh(2m\beta)}{(2\sinh\beta)^{2m-1}\cosh\beta}\right)^{1/m} \quad (120)$$

$$= \lim_{m\to\infty}\left(\frac{1}{4m^2\sinh^2\beta}\right)\left(\frac{2\sin\beta\sinh(2m\beta)}{\cosh\beta}\right)^{1/m} \quad (121)$$

$$= \lim_{m\to\infty}\frac{K(m,\gamma)}{2m^2} = \frac{1}{(\cosh^{-1}\gamma)^2}, \quad (122)$$

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

Motivated by the above proposition, let us define

$$w_j := 1 + \sigma j^2, \qquad j \geq 1 \ldots, \tag{123}$$

and an associated vector sequence $\mathbf{w}^{(m)} := (w_1, \ldots, w_{m-1})$, $m \geq 2$. We know that for each $j \geq 1$, $x_j^{(m)} \to w_j$ as $m \to \infty$. The following lemma provides a nonasymptotic relation between $\mathbf{x}^{(m)}$ and $\mathbf{w}^{(m)}$ which will be utilized in proving Proposition 5.3 and Theorem 5.5.

Lemma 5.2. $\mathbf{w}^{(m)}$ *is subordinate to* $\mathbf{x}^{(m)}$.

Proof. By Definition 3.2, we need to show that for $0 \leq j \leq m-2$ and $w_0^{(m)} \equiv x_0^{(m)} \equiv 1$,

$$\frac{w_{j+1}^{(m)}}{x_{j+1}^{(m)}} \geq \frac{w_j^{(m)}}{x_j^{(m)}}, \tag{124}$$

that is,

$$(1+\sigma(j+1)^2)\left(1 + 2K\sin^2\left(\frac{j\pi}{2m}\right)\right) \geq (1+\sigma j^2)\left(1 + 2K\sin^2\left(\frac{(j+1)\pi}{2m}\right)\right), \tag{125}$$

or equivalently

$$\begin{aligned} &\left[\sigma(2j+1) - 2K\left(\sin^2\left(\frac{(j+1)\pi}{2m}\right) - \sin^2\left(\frac{j\pi}{2m}\right)\right)\right] \\ &+ \left[2\sigma K\left((j+1)^2\sin^2\left(\frac{j\pi}{2m}\right) - j^2\sin^2\left(\frac{(j+1)\pi}{2m}\right)\right)\right] \;\geq\; 0. \end{aligned} \tag{126}$$

We show that both these summands are nonnegative.

By Proposition 5.1(a), $K \leq \frac{2m^2\sigma}{\pi^2}$, and so for the first summand, it is sufficient to show that

$$(2j+1) - \frac{4m^2}{\pi^2}\left(\sin^2\left(\frac{(j+1)\pi}{2m}\right) - \sin^2\left(\frac{j\pi}{2m}\right)\right) \geq 0. \tag{127}$$

Indeed, by standard trigonometric identities

$$\frac{4m^2}{\pi^2}\left(\sin^2\left(\frac{(j+1)\pi}{2m}\right) - \sin^2\left(\frac{j\pi}{2m}\right)\right) = \frac{4m^2}{\pi^2}\sin\left(\frac{(2j+1)\pi}{2m}\right)\sin\left(\frac{\pi}{2m}\right) \leq 2j+1, \tag{128}$$

which proves (127).

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 (126) and hence the proof of the Lemma. □

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 3.2 to $\mathbf{x}^{(m)} := (1 + K(1+z_j))_{j=1}^{m-1}$, the solution

of the relaxed minimization problem (41), (42). By Corollary 3.3, $f(\mathbf{n}^{(m)}) \le f(\mathbf{x}^{(m)}) \le \gamma$, so $\mathbf{n}^{(m)}$ satisfies (40). Note that for the $n_j^{(m)}$'s we use the original index set $j = 1, \dots, m$ of Section 2.4, while for the $x_j^{(m)}$'s we retain the labels $j = 0, \dots, m-1$. In this section, we work with the specific integer sequence $\mathbf{n}^{(m)} = (n_2^{(m)}, \dots, n_m^{(m)})$ defined recursively by

$$n_{j+1}^{(m)} = \left\lceil n_j^{(m)} \frac{x_j^{(m)}}{x_{j-1}^{(m)}} \right\rceil, \qquad j = 1, \dots, m-1, \tag{129}$$

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 $\mathbf{k} = (k_2, \dots, k_m)$ is any integer sequence such that $1 \le \frac{k_2}{x_1^{(m)}} \le \cdots \le \frac{k_m}{x_{m-1}^{(m)}}$, then $k_j \ge n_j^{(m)}$ for all $j = 2, \dots, m$. Indeed one has $k_2 \ge x_1^{(m)}$, which implies $k_2 \ge \lceil x_1^{(m)} \rceil = n_2^{(m)}$, and assuming by induction $k_j \ge n_j^{(m)}$, one obtains

$$k_{j+1} = \lceil k_{j+1} \rceil \ge \left\lceil x_j^{(m)} \frac{k_j}{x_{j-1}^{(m)}} \right\rceil \ge \left\lceil x_j^{(m)} \frac{n_j^{(m)}}{x_{j-1}^{(m)}} \right\rceil = n_{j+1}^{(m)}. \tag{130}$$

We will next derive an asymptotic formula, analogous to Proposition 5.1(c), for the limit of $n_j^{(m)}$ as $m \to \infty$.

Proposition 5.3. *Let $\sigma > \frac{5}{4}$. Then, for all $j \ge 1$,*

$$\lim_{m \to \infty} n_j^{(m)} = 1 + \lceil \sigma \rceil (j-1)^2. \tag{131}$$

Proof. We will prove the statement by induction on j. The case $j = 1$ holds by definition. Assume that (131) holds for some $j \ge 1$.

We first consider the case when σ is an integer. Since $n_j^{(m)}$ is integer valued and $\lceil \sigma \rceil = \sigma$, (131) is equivalent to saying that $n_j^{(m)} = 1 + \sigma(j-1)^2$ for all sufficiently large m. Now by Lemma 5.2 (see (124)) we have

$$1 + \sigma j^2 \ge (1 + \sigma(j-1)^2) \frac{x_j^{(m)}}{x_{j-1}^{(m)}} = n_j^{(m)} \frac{x_j^{(m)}}{x_{j-1}^{(m)}} \tag{132}$$

which, together with (129), implies $1 + \sigma j^2 \ge n_{j+1}^{(m)}$. Since $n_{j+1}^{(m)} \ge x_j^{(m)}$ and $x_j^{(m)} \to 1 + \sigma j^2$ (see Proposition 5.1(c)), we obtain $n_{j+1}^{(m)} \to 1 + \sigma j^2$, which completes the induction step.

Now assume $\sigma > \frac{5}{4}$ is noninteger. By Proposition 5.1(c) and the induction hypothesis, we have

$$\lim_{m \to \infty} n_j^{(m)} \frac{x_j^{(m)}}{x_{j-1}^{(m)}} = (1 + \lceil \sigma \rceil (j-1)^2) \frac{1 + \sigma j^2}{1 + \sigma(j-1)^2} \tag{133}$$

Since the function $\lceil\cdot\rceil$ is continuous at every noninteger, our problem reduces to showing that

$$\lceil\sigma\rceil j^2 < (1+\lceil\sigma\rceil(j-1)^2)\frac{1+\sigma j^2}{1+\sigma(j-1)^2} < 1+\lceil\sigma\rceil j^2 \tag{134}$$

for all $j \geq 1$ and $\sigma > \frac{5}{4}$, for then we have

$$1+\lceil\sigma\rceil j^2 = \left\lceil \lim_{m\to\infty} n_j^{(m)} \frac{x_j^{(m)}}{x_{j-1}^{(m)}} \right\rceil = \lim_{m\to\infty} \left\lceil n_j^{(m)} \frac{x_j^{(m)}}{x_{j-1}^{(m)}} \right\rceil = \lim_{m\to\infty} n_{j+1}^{(m)}, \tag{135}$$

which would complete the induction step.

To show (134), first note that

$$(1+\lceil\sigma\rceil(j-1)^2)\frac{1+\sigma j^2}{1+\sigma(j-1)^2} - \lceil\sigma\rceil j^2 = 1 - \frac{(\lceil\sigma\rceil-\sigma)(2j-1)}{1+\sigma(j-1)^2}, \tag{136}$$

which immediately yields the second inequality for all $j \geq 1$ and noninteger $\sigma > 0$. The first inequality in (134) is also immediate for $j = 1$, as the right hand side of (136) reduces to $1-(\lceil\sigma\rceil-\sigma)$, which is strictly positive. Now note that

$$1+\sigma(t-1)^2 \geq 2\sigma t + 1 - 3\sigma, \text{ for all } t \in \mathbb{R}, \tag{137}$$

since the right hand side is the equation of the line tangent to the parabola $1+\sigma(t-1)^2$ at $t = 2$. Now, for $\sigma > 2$, we have

$$\frac{(\lceil\sigma\rceil-\sigma)(2j-1)}{1+\sigma(j-1)^2} < \frac{2j-1}{1+2(j-1)^2} \leq \frac{2j-1}{4j-5} \leq 1, \text{ for all } j \geq 2. \tag{138}$$

On the other hand, for $\frac{5}{4} < \sigma < 2$, we have

$$\frac{(\lceil\sigma\rceil-\sigma)(2j-1)}{1+\sigma(j-1)^2} < \frac{\frac{3}{4}(2j-1)}{1+\frac{5}{4}(j-1)^2} \leq \frac{\frac{3}{2}j-\frac{3}{4}}{\frac{5}{2}j-\frac{11}{4}} \leq 1, \text{ for all } j \geq 2. \tag{139}$$

Hence the first inequality in (134) holds for all $j \geq 1$ and noninteger $\sigma > \frac{5}{4}$. □

Remark: For $j = 1$ and $j = 2$, we have $n_1^{(m)} = 1$ and $n_2^{(m)} = \lceil x_1^{(m)}\rceil$ so that the formula (131) is actually valid for all $\sigma > 0$ because of Proposition 5.1(b) and (c). On the other hand, for $j = 3$ and $1 < \sigma < \frac{5}{4}$, we have $n_3^{(m)} \to 8$ instead of 9. To see this, note that $n_2^{(m)} \to 1+\lceil\sigma\rceil = 3$, so that

$$\lim_{m\to\infty} n_2^{(m)} \frac{x_2^{(m)}}{x_1^{(m)}} = \frac{3(1+4\sigma)}{1+\sigma} \in (7.5, 8). \tag{140}$$

Similarly, for $1 < \sigma < \frac{5}{4}$, $n_4^{(m)} \to 17$ instead of 19. As shown in Figure 1, the pattern becomes more complicated for higher values of j and the interval has to be subdivided. A similar pattern is present also for $0 < \sigma < 1$. We shall not explore this here; we again refer to Figure 1.

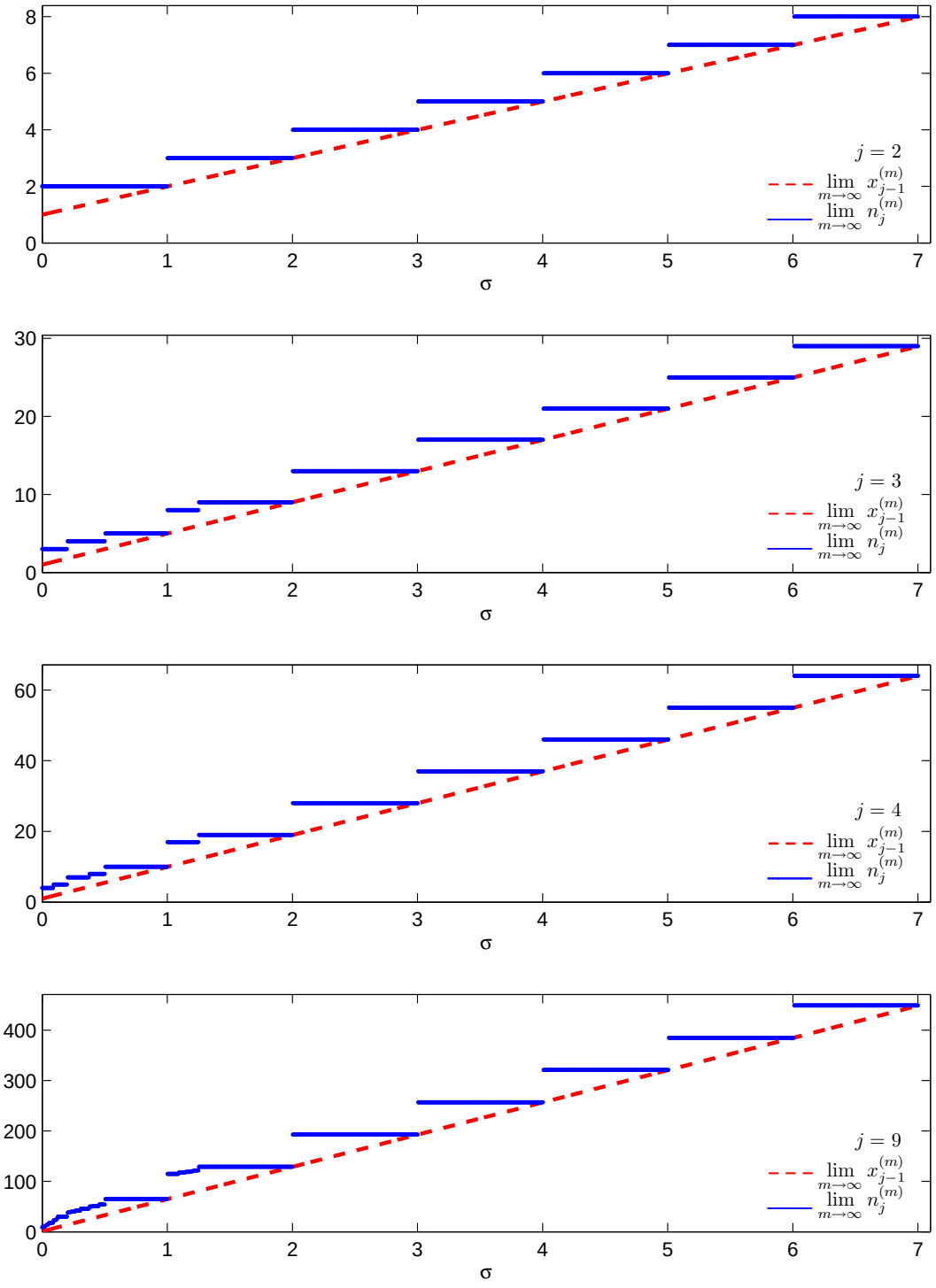

Figure 1: Numerical comparison of $\lim_{m\to\infty} x_{j-1}^{(m)}$ with $\lim_{m\to\infty} n_j^{(m)}$ as a function of σ for $j = 2, 3, 4, 9$. For $j = 2$, these limits equal $1+\sigma$ and $1+\lceil\sigma\rceil$, respectively. For $j \geq 3$, notice the difference between the cases $\sigma > 5/4$ and $\sigma < 5/4$. In the first case, one has $\lim_{m\to\infty} n_j^{(m)} = 1 + \lceil\sigma\rceil(j-1)^2$, whereas in the latter case this formula is no longer valid.

Definition 5.4. *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*

$$\lim_{m\to\infty} \left(\frac{\eta\left(\mathbf{k}^{(m)}\right)}{\eta\left(\mathbf{x}^{(m)}\right)} \right)^{1/m} = 1, \tag{141}$$

where $\mathbf{x}^{(m)}$ *is the solution of (41), (42) as above.*

The relevance of this definition lies in the fact that the quantity $\lim_{m\to\infty} \frac{1}{m^2}(\eta(\mathbf{k}^{(m)}))^{1/m}$ controls the rate of exponential decay for the error bound associated with the sequence of filters defined by $\left(\mathbf{k}^{(m)}\right)$. The precise relation will be seen in Theorem 5.6 below. We will use the following lemma to assess the asymptotic optimality of our minimal subordinate construction $\mathbf{n}^{(m)}$ defined in (129).

Theorem 5.5. *If* σ, *defined in Proposition 5.1, is an integer, then the sequence* $\mathbf{n}^{(m)}$, $m = 2, 3, \ldots$, *defined by (129), is both asymptotically optimal and subordinate to* $\mathbf{x}^{(m)}$. *If* σ *is not an integer, no sequence of integer vectors can have both of 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 $\mathbf{x}^{(m)}$,

$$\limsup_{m\to\infty} \left(\frac{\eta(\mathbf{k})}{\eta(\mathbf{x})} \right)^{1/m} \geq \limsup_{m\to\infty} \frac{k_2^{(m)}}{x_1^{(m)}} \geq \lim_{m\to\infty} \frac{\lceil x_1^{(m)} \rceil}{x_1^{(m)}} = \frac{1+\lceil\sigma\rceil}{1+\sigma} > 1. \tag{142}$$

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

Now consider the case that σ is an integer. Then by Lemma 5.2, $\mathbf{w}^{(m)} = (w_1, \ldots, w_{m-1})$ is an integer sequence subordinate to $\mathbf{x}^{(m)}$. As in Proposition 5.3, one has $x_j^{(m)} \leq n_{j+1}^{(m)} \leq w_j$, as $\mathbf{n}^{(m)}$ is the minimal integer sequence subordinate to $\mathbf{x}^{(m)}$.

Next, from Proposition 5.1(a) we see that

$$\frac{2K}{m^2} \geq \frac{4s}{\pi^2}\left(1 - \frac{C_1}{m}\right) \tag{143}$$

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 $C_3 < \infty$

$$x_j^{(m)} = 1+2K\sin^2\left(\frac{j\pi}{2m}\right) \geq 1+\sigma\left(1-\frac{C_1}{m}\right)\left(j^2\left(1 - C_2\frac{\pi^2}{4}\frac{j^2}{m^2}\right)\right) \geq (1+\sigma j^2)\left(1-\frac{C_3}{m^{2/3}}\right). \tag{144}$$

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

$$\frac{{n_{j+1}}^{(m)}}{{x_j}^{(m)}} \leq \frac{w_j}{{x_j}^{(m)}} \leq \frac{1}{1-\frac{C_3}{m^{2/3}}} \leq 1 + \frac{C_4}{m^{2/3}}. \tag{145}$$

Now

$$\frac{{n_{j+1}}^{(m)}}{{x_j}^{(m)}} = \frac{1}{{x_j}^{(m)}} \left\lceil n_j^{(m)} \frac{x_j^{(m)}}{x_{j-1}^{(m)}} \right\rceil \leq \frac{1}{{x_j}^{(m)}} \left(n_j^{(m)} \frac{x_j^{(m)}}{x_{j-1}^{(m)}} + 1 \right) = \frac{n_j^{(m)}}{x_{j-1}^{(m)}} + \frac{1}{{x_j}^{(m)}}. \tag{146}$$

Combining (143) 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 (146), one then obtains for $m^{2/3} < j \leq m-1$ and some constant $C_6 < \infty$

$$\frac{n_{j+1}^{(m)}}{x_j^{(m)}} \leq \frac{n_j^{(m)}}{x_{j-1}^{(m)}} + \frac{1}{C_5 j^2} \leq \cdots \leq \frac{n_{\lfloor m^{2/3} \rfloor}^{(m)} + 1}{x_{\lfloor m^{2/3} \rfloor}^{(m)}} + \sum_{l=\lfloor m^{2/3} \rfloor + 1}^{j} \frac{1}{C_5 l^2} \leq 1 + \frac{C_4}{m^{2/3}} + \frac{C_6}{m^{2/3}}. \tag{147}$$

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

$$\frac{{n_{j+1}}^{(m)}}{{x_j}^{(m)}} \leq 1 + \frac{C_7}{m^{2/3}}. \tag{148}$$

We conclude that

$$1 \leq \frac{\eta(\mathbf{n}^{(m)})}{\eta(\mathbf{x}^{(m)})} = \prod_{j=1}^{m-1} \frac{n_{j+1}^{(m)}}{{x_j}^{(m)}} \leq \left(1 + \frac{C_7}{m^{2/3}} \right)^m, \tag{149}$$

which implies that

$$\lim_{m \to \infty} \left(\frac{\eta(\mathbf{n}^{(m)})}{\eta(\mathbf{x}^{(m)})} \right)^{1/m} = 1, \tag{150}$$

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

Remark: For noninteger values of σ, we do not know if there are asymptotically optimal integer vectors $\mathbf{k}^{(m)}$ (which are necessarily not subordinate to $\mathbf{x}^{(m)}$) that are admissible, i.e., that satisfy $f(\mathbf{k}^{(m)}) \leq \gamma = \cosh(\pi/\sqrt{\sigma})$. At the same time, it is natural to ask how close our construction $\mathbf{n}^{(m)}$ is to being asymptotically optimal, i.e., the value of the limit, as $m \to \infty$, of $(\eta(\mathbf{n}^{(m)})/\eta(\mathbf{x}^{(m)}))^{1/m}$. We shall not carry out an analysis here that is similar to Proposition 5.3 to find this limit, but instead provide our numerical findings in Figure 2.

Optimal exponential error decay for minimally supported filters

We are now ready to prove the promised improved exponential error decay estimate for the $\Sigma\Delta$ modulators defined by $\mathbf{n}^{(m)}$.

Theorem 5.6. *For all $1 < \gamma < 2$ such that $\sigma = \frac{\pi^2}{\left(\cosh^{-1}\gamma\right)^2}$ is an integer, all one-bit $\Sigma\Delta$ 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 $\|y\|_\infty \leq \mu = 2 - \gamma$. Furthermore, the family consisting of*

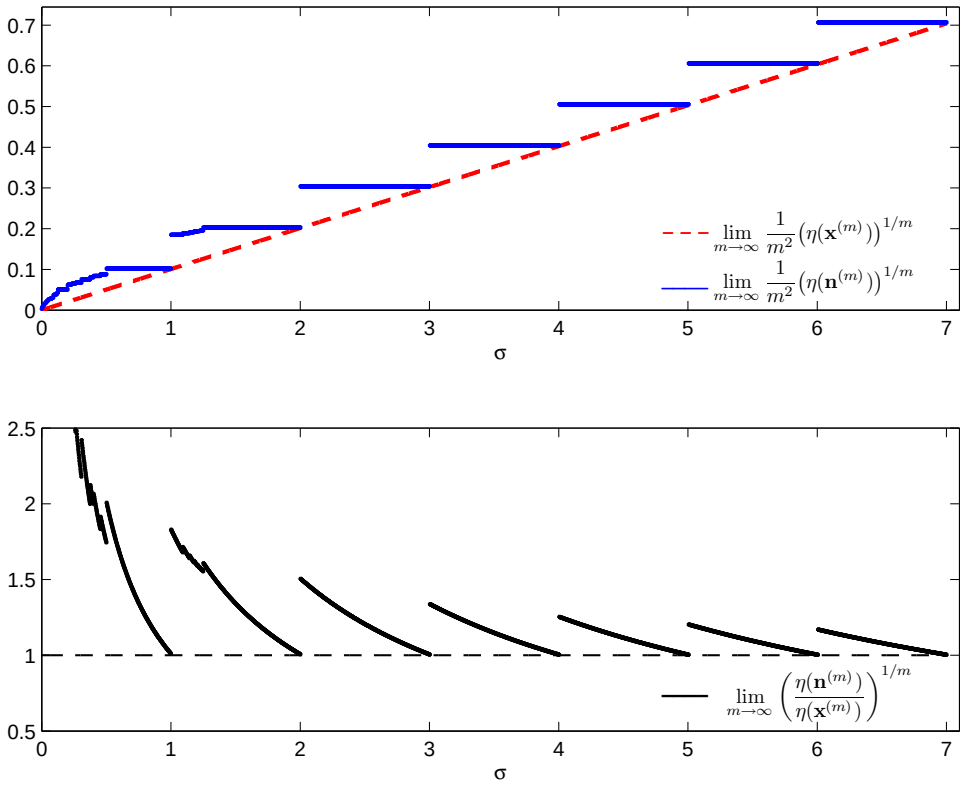

Figure 2: Numerical comparison of $\lim_{m\to\infty} \frac{1}{m^2}\left(\eta(\mathbf{n}^{(m)})\right)^{1/m}$ with $\lim_{m\to\infty} \frac{1}{m^2}\left(\eta(\mathbf{x}^{(m)})\right)^{1/m} = \sigma/\pi^2$, as a function of σ. Top: plotted individually, bottom: their ratio.

the one-bit $\Sigma\Delta$ 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

$$\|e_\lambda\|_\infty \leq C 2^{-r\lambda}. \tag{151}$$

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_0 such that the corresponding ϵ as introduced in Section 2.2 satisfies $1 + \epsilon < \sqrt{\frac{r_0}{r}}$. Now let $g^{(m)}$ be such that $\Delta^m g^{(m)} = \delta^{(0)} - h^{(m)}$, as in Section 2.3. Then from (18) and (27), we have the error bound

$$\|e_\lambda\|_\infty \leq \|g^{(m)}\|_1 \|v\|_\infty \|\varphi_0\|_1 \pi^m (1+\epsilon)^m \lambda^{-m}, \tag{152}$$

where v solves (26).

Recall that our construction yields $\|v\|_\infty \leq 1$. Furthermore, by Theorem 5.5 and Proposition 5.1 (d), we have that

$$\lim_{m\to\infty} \frac{\left(\eta(\mathbf{n}^{(m)})\right)^{1/m}}{m^2} = \frac{1}{(\cosh^{-1}\gamma)^2} = \frac{\sigma}{\pi^2} \tag{153}$$

and hence by (34)

$$\|g^{(m)}\|_1 = \frac{\eta(\mathbf{n}^{(m)})}{m!} = \left(\frac{e\sigma}{\pi^2}\right)^m m^m \left(1 + o(1)\right)^m. \tag{154}$$

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 (152) implies

$$\|e_\lambda\|_\infty \leq \|\varphi_0\|_1 \left(\frac{e\sigma}{\pi}\right)^m m^m \left(\frac{r_0}{r}\right)^m \lambda^{-m}. \tag{155}$$

As explained in Section 2.3, we choose, for each λ, the filter $h^{(m)}$ that leads to the minimal error bound. Then by a slight variation of (23), we obtain

$$\|e_\lambda\|_\infty \leq \|\varphi_0\|_1 \min_{m \geq M(r)} \left(\frac{e\sigma}{\pi}\right)^m m^m \left(\frac{r_0}{r}\right)^m \lambda^{-m} \lesssim_r \exp\left(-\frac{\pi}{e^2\sigma}\frac{r}{r_0}\lambda\right) = 2^{-r\lambda}, \tag{156}$$

which proves the theorem. □

Remark: The smallest integer σ such that the stability constraint $\|h\|_1 \leq \gamma$ is satisfied for some $\gamma < 2$ is $\sigma = 6$. In this case, Theorem 5.6 yields exponential error decay for any rate constant $r < r_0 \approx 0.102$. This is the fastest error decay currently known to be achievable for one-bit $\Sigma\Delta$ modulation. The previously best known bound for the achievable rate constant was $r_0 \approx 0.088$ [11].

6 Multi-level quantization alphabets and the case of small σ

In this paper we have primarily considered one-bit $\Sigma\Delta$ modulators, though our theory and analysis is equally applicable to quantization alphabets that consist of more than two levels. Let us consider a general alphabet $\mathcal{A} = \mathcal{A}_L$ with L levels such that consecutive levels are separated by 2 units as before. For instance, $\mathcal{A}_4 = \{-3, -1, 1, 3\}$ and $\mathcal{A}_5 = \{-4, -2, 0, 2, 4\}$. The corresponding generalized stability condition for the greedy quantization rule is then

$$\|h\|_1 + \|y\|_\infty \leq L, \tag{157}$$

which still guarantees the bound $\|v\|_\infty \leq 1$ (see, e.g., [13, p. 104]). In terms of our filter design and optimization problem, the parameter γ that bounds $\|h\|_1$ can be set as large as L. The potential benefit of increased number of levels is that large values of γ correspond to small values of σ, which in turn yields faster exponential error decay rates in the oversampling ratio λ as given in (156). There is, however, also an increase in the number of bits spent per sample which needs to be accounted for if a rate-distortion type performance analysis is to be carried out.

As seen above (Theorem 5.5 and Figure 2), our analytical results are available for integer values of σ only. We will evaluate the performance of multi-level quantization alphabets

L	2	3	4	5	12
(average) number of bits/sample $B = \log_2 L$	1	1.585	2	2.322	3.585
minimum integer value of σ	6	4	3	2	1
maximum input signal $\|y\|_\infty$	0.058	0.490	0.851	0.335	0.408
achievable error decay rate $r_0 = \pi(e^2\sigma \ln 2)^{-1}$	0.102	0.153	0.204	0.306	0.613
coding efficiency r_0/B	0.102	0.097	0.102	0.132	0.171

Table 1: Comparison of exponential error decay rates and their coding efficiency for multi-level quantization alphabets.

based on these values first. For each positive integer σ, we can find the minimum integer value of $L \geq 2$ such that $\cosh(\pi/\sqrt{\sigma}) < L$. Alternatively, if $L \geq 2$ is specified first, we find the minimum integer value of σ that satisfies this condition. Given such a (σ, L) pair, the corresponding bound $\|y\|_\infty$ on the input signal is $L - \cosh(\pi/\sqrt{\sigma})$. We also compute the achievable exponential error decay rate r_0 given by $\frac{\pi}{e^2\sigma \ln 2}$, the average number of quantizer bits per sample $B := \log_2 L$, and a coding efficiency figure given by r_0/B. The results are tabulated in Table 1. It turns out that the smallest value of L which results in $\sigma = 1$ is $L = 12$. This case also yields the highest coding efficiency figure given by 0.171. It also yields a significantly more favorable range for the input signal compared to the one-bit case ($L = 2$). As before, it is easy to check via Kolmogorov entropy bounds that the coding efficiency is always bounded by 1. The cases $L = 4$ and $L = 5$ are also noteworthy for they provide better overall performance, yet still with a small quantization alphabet.

It is natural to ask what can be said for the case of noninteger values of σ, especially as $\sigma \to 0$. For our minimal subordinate constructions, we employ the numerically computed values of $\lim_{m\to\infty} \frac{1}{m^2}\big(\eta(\mathbf{n}^{(m)})\big)^{1/m}$ given in Figure 2 as a function of the continuous parameter σ. As in Theorem 5.6, the achievable error decay rate r_0 is given by

$$r_0 = \frac{1}{\pi e^2 \ln 2} \lim_{m\to\infty} \frac{m^2}{\big(\eta(\mathbf{n}^{(m)})\big)^{1/m}}. \tag{158}$$

On the other hand the smallest number of levels L of the quantizer which guarantees stability for the greedy rule is given by

$$L := \lceil \cosh(\pi/\sqrt{\sigma}) \rceil. \tag{159}$$

Finally, the coding efficiency as a function of σ is

$$\frac{r_0}{B} = \frac{1}{\pi e^2 \ln\lceil \cosh(\pi/\sqrt{\sigma})\rceil} \lim_{m\to\infty} \frac{m^2}{\big(\eta(\mathbf{n}^{(m)})\big)^{1/m}}. \tag{160}$$

We plot our numerical findings in Figure 3. The specific values reported in Table 1 are visible at the integer values $\sigma = 1, 2, 3, 4, 6$. It is interesting that the coding efficiency figure 0.171 reported for $\sigma = 1$ seems to be the global maximum value achievable by our construction over all values of σ. This is possibly the best performance achievable by minimally supported

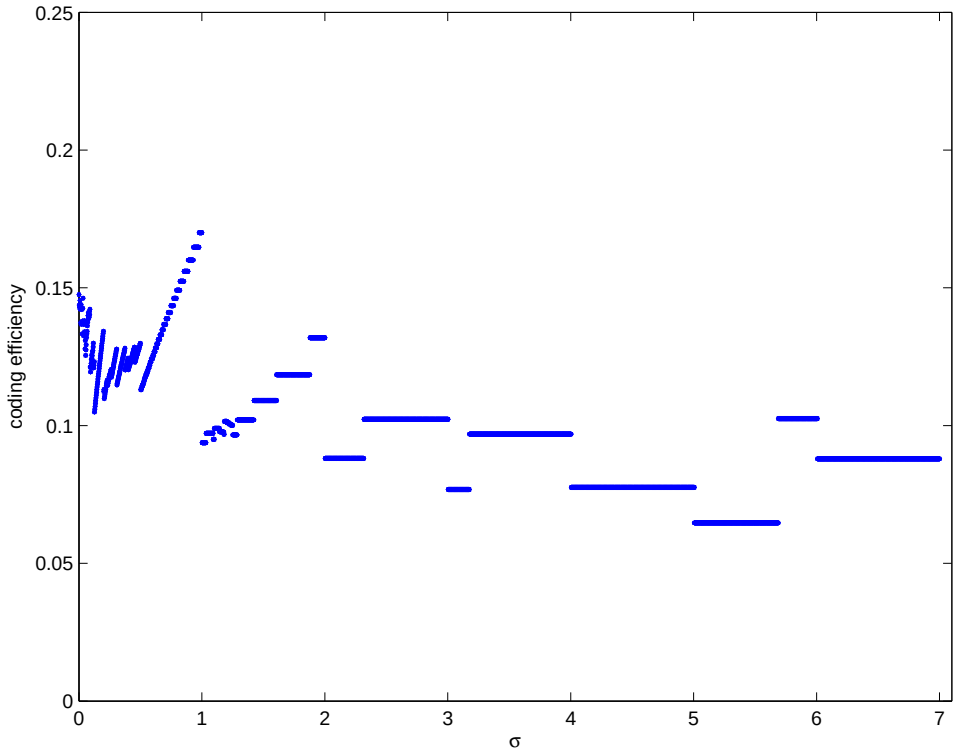

Figure 3: Numerical computation of the coding efficiency formula given by (160) for the minimal subordinate filters $\mathbf{n}^{(m)}$.

filters that are constrained by subordinacy. We do not know if this figure can be exceeded without the subordinacy constraint. The case of arbitrary optimal filters (i.e., not necessarily minimally supported) is also open, along with the case of quantization rules that are more general than the greedy rule.

Acknowledgements

The work in this paper was supported in part by various grants and fellowships: National Science Foundation Grants DMJ-0500923 (Deift), CCF-0515187 (Güntürk), Alfred P. Sloan Research Fellowship (Güntürk), the Morawetz Fellowship at the Courant Institute (Krahmer), the Charles M. Newman Fellowship at the Courant Institute (Krahmer) and a NYU GSAS Dean's Student Travel Grant (Krahmer).

A Some useful properties of Chebyshev polynomials

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

$$T_m(x) = \cos m\theta \qquad \text{and} \qquad U_m(x) = \frac{\sin(m+1)\theta}{\sin\theta}, \tag{161}$$

respectively. The Chebyshev polynomials have, in particular, the following properties (see [14], [3]):

- $T_m'(x) = mU_{m-1}(x)$,
- 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

$$T_m(\cosh\tau) = \cosh(m\tau), \quad U_m(\cosh\tau) = \frac{\sinh(m\tau)}{\sinh\tau}, \tag{162}$$

- The Chebyshev polynomials satisfy the differential equation

$$(1-x)^2 T_m''(x) - xT_m'(x) + m^2 T_m(x) = 0. \tag{163}$$

We say that a polynomial p of degree m has the *equi-oscillation property* on $[-1, 1]$ (compare [3]) if it has $m-1$ real critical points $\zeta_1, \ldots, \zeta_{m-1}$ which satisfy

$$\zeta_0 := -1 < \zeta_1 < \cdots < \zeta_{m-1} < \zeta_m := 1 \tag{164}$$

such that the associated values are alternating

$$p(\zeta_j) = (-1)^{m-j} \tag{165}$$

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 A.1. *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 [3] to establish that, up to a constant, the T_m are the unique monic polynomials with minimal L^∞ norm.

Let $p(s) = a_p s^m + \dots$ and $q(s) = a_q s^m + \dots$ be two polynomials with the equi-oscillation property. W.l.o.g. assume $a_q \geq a_p > 0$. Let $\zeta_1 < \dots < \zeta_{m-1}$ 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 $t_0 < t_1 < \dots < t_m \in \mathbb{R}$ and a polynomial ρ of degree $m-1$ satisfies $(-1)^j \rho(\zeta_j) \geq 0$ for all j, then $\rho \equiv 0$.*

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 $p \equiv q$. □

Proof of Claim: 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 $\tilde{\rho} \equiv 0$. □

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

Lemma A.2. *Let z_j, $j = 1, \dots, m-1$, be the critical points of the Chebyshev polynomial of the first kind T_m, as above, and set $z_0 \equiv -1$. Then*

$$\prod_{\substack{i=0 \\ i \neq k}}^{m-1} (z_k - z_i) = \begin{cases} \frac{m(-1)^{m-1}}{2^{m-1}} & \text{for } k = 0 \\ \\ \frac{m(-1)^{m-1-k}}{2^{m-1}(1-z_k)} & \text{for } k > 0 \end{cases} \tag{166}$$

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

$$T_m'(z) = m 2^{m-1} \prod_{i=1}^{m-1} (z - z_i), \tag{167}$$

and

$$T_m''(z) = m 2^{m-1} \sum_{j=1}^{m-1} \prod_{\substack{i=1 \\ i \neq j}}^{m-1} (z - z_i), \tag{168}$$

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

$$T_m''(z_k) = m2^{m-1} \prod_{\substack{i=1 \\ i \neq k}}^{m-1} (z_k - z_i) \tag{169}$$

$$= \frac{m2^{m-1}}{1+z_k} \prod_{\substack{i=0 \\ i \neq k}}^{m-1} (z_k - z_i). \tag{170}$$

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 (163) reads

$$(1-z_k^2)\frac{m2^{m-1}}{1+z_k} \prod_{\substack{i=0 \\ i \neq k}}^{m-1} (z_k - z_i) + m^2(-1)^{m-k} = 0, \tag{171}$$

or

$$\prod_{\substack{i=0 \\ i \neq k}}^{m-1} (z_k - z_i) = \frac{(-1)^{m-k-1}m}{2^{m-1}(1-z_k)}. \tag{172}$$

On the other hand, as $z_0 = \cos(\pi)$, we have using (161)

$$\prod_{i=1}^{m-1} (z_0 - z_i) = \frac{1}{m2^{m-1}} T_m'(z_0) = \frac{1}{2^{m-1}} U_m(z_0) = \frac{1}{2^{m-1}} \lim_{\theta \to \pi} \frac{\sin(m\theta)}{\sin\theta} = \frac{(-1)^{m-1}m}{2^{m-1}}. \tag{173}$$

□

References

[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] B. Bodmann and V. I. Paulsen. Frame paths and error bounds for sigma-delta quantization. *Appl. Comput. Harmon. Anal.*, 22, 2007.

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

[4] 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.

[5] 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.

[6] I. Daubechies, R. DeVore, C. S. Güntürk, and V. A. Vaishampayan. A/D conversion with imperfect quantizers. *IEEE Trans. Inform. Theory*, 52(3):874–885, 2006.

[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] H. Inose, Y. Yasuda, and J. Murakami. A telemetering system by code manipulation - $\Delta\Sigma$ modulation. *IRE Trans on Space Electronics and Telemetry*, pages 204–209, 1962.

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

[13] R. Schreier and G. C. Temes. *Understanding Delta-Sigma Data Converters*. Wiley-IEEE Press, 2004.

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

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