On quantization of finite frame expansions: sigma-delta schemes of arbitrary order

M.C. Lammers [a], Alexander M. Powell [b] and Özgür Yılmaz [c]

[a] Dept. of Mathematics, University of North Carolina Wilmington, NC, 28403
[b]Vanderbilt University, Department of Mathematics, Nashville, TN 37240
[c] Department of Mathematics, University of British Columbia, Vancouver, B.C. Canada V6T 1Z2

ABSTRACT

In this note we will show that the so called Sobolev dual is the minimizer over all linear reconstructions using dual frames for stable r^{th} order $\Sigma\Delta$ quantization schemes under the so called White Noise Hypothesis (WNH) design criteria. We compute some Sobolev duals for common frames and apply them to audio clips to test their performance against canonical duals and another alternate dual corresponding to the well known Blackman filter.

Keywords: quantization, sigma-delta, alternate duals, Sobolev dual

1. INTRODUCTION

Many digital signal representations take advantage of oversampling the signal, i.e., these representations use excess information about the signal to offset information lost when converting from an analog setting to a digital one. Recently frames have been used to describe *analog-to-digital (AD) conversion* and specifically the quantization algorithm known as $\Sigma\Delta$. While there are numerous papers on $\Sigma\Delta$ in the engineering literature,[4,8,10,11,15,19,20] to which this list does not begin to do justice, a paper of Daubechies and DeVore [9] created great interest in the mathematical community [12–14,17,21]. In particular, Benedetto, Powell and Yilmaz [1,2] recently used finite frames to describe and examine first order $\Sigma\Delta$ schemes, followed by Bodmann and Paulsen [6]. Of greatest interest in these papers is the convergence rates of the algorithm and convergence constants associated with different types of frames.

As we will describe later, $\Sigma\Delta$ has a number of different variations, which we will refer to as schemes, that correspond to the number of differences applied in the algorithm. Higher order schemes in the finite frame setting have also been previously studied [3] and continue to be examined [5,7,18]. The higher order $\Sigma\Delta$ schemes for finite frames have provided a few surprises in that some of the early convergence rates for the group of frames being studies were somewhat unexpected given results from the band limited case. This is usually explained by the fact that there are so called "boundary terms" in the finite sums of the reconstruction that do not occur in the infinite, band limited case. Two approaches have been developed to compensate for these boundary terms and hence achieve the desired rates of convergence for a scheme of a particular order. One approach is to chose a different class of frames for which to perform the sigma delta quantization on and the second is to consider alternate duals for reconstruction. This paper focusses on the later.

The main goal of this note is to examine the MSE of the reconstruction of higher order $\Sigma\Delta$ schemes with the so called *Sobolev dual* introduced in [5]. Although it is well known to engineers that the so called *White Noise Hypothesis(WNH)* does not hold in general, it has been successfully used as a design criteria for many years. For some reasons why this is true we refer the reader to an article by Gray [16]. In what follows, we will show that the so called Sobolev dual is optimal for minimizing the*Mean Square Error (MSE)* under the WNH and test it against some other dual frames.

E-mail:lammersm@uncw.edu, alexander.m.powell@vanderbilt.edu, oyilmaz@math.ubc.ca

2. FRAMES

DEFINITION 2.1. *Let $\mathcal{H}$ be a Hilbert space, then a sequence of elements $\{g_n\}$ is called a* **frame** *if there exist A, B with $0 < A < B < \infty$ so that for all $f \in \mathbb{H}$*

$$A\|f\|^2 \leq \sum_n |\langle f, g_n\rangle|^2 \leq B\|f\|^2.$$

The operator $S_g(f) = \sum_n \langle f, g_n\rangle g_n$ is called the **frame operator** *and is a self adjoint invertible operator.*

Any sequence $\{h_n\}$ so that $f = \sum_n \langle f, g_n\rangle h_n$ is called a **dual frame** *of $\{g_n\}$ and the sequence given by $\{S_g^{-1}(g_n)\}$ is a dual frame and it is referred to as the* **canonical dual**.

If $A = B$ the frame is called **tight** and the canonical dual is $\frac{1}{A}\{g_n\}$.

EXAMPLE 2.2.

1) The rows of any matrix E^*, even non-square, so that EE^*, i.e., the frame operator, is invertible.

2) Frame Paths: let $E(t) : [0,1] \to \mathbb{R}^d$ where $E^*(t) = [e_1(t), e_2(t), \cdots e_d(t)]$. $E^*(t)$ is a **frame path** *(see[6])if the rows of the matrix obtained by uniform sampling $[E^*(\frac{i}{N})]_{N\times d}$ is a frame for $\mathbb{R}^d$*

a) **Harmonic frame path** *[10,11,22]*

$E^*(t) = [\cos(2\pi t), \sin(2\pi t), \cdots \cos(2k\pi t), \sin(2k\pi t)]$ *and* $d = 2k$

b) **Sampling frame path***(repeated o.n basis) for $\mathbb{R}^d$*

$E^*(t) = [\lambda_{[0,\frac{1}{d})}(t), \lambda_{[\frac{1}{d},\frac{2}{d})}(t), \cdots \lambda_{[\frac{d-1}{d},1)}(t)]$, *where λ is the indicator function.*

3. QUANTIZATION.

Given an expansion of the form $f(x) = \sum_n a_n e_n$, find an expression $\tilde{f}(x) = \sum_n q_n e_n$so $\|f - \tilde{f}\|$is *small* **and** q_n come from a finite alphabet. For example , $q_n \in \{-1, 1\}$.

EXAMPLE 3.1. *For* **bandlimited functions** *$f(x)$ with $|\hat{f}(w)|, |f(x)| < 1$, having the representation $f(x) = \frac{1}{\lambda}\sum_n f(\frac{n}{\lambda})g(x - \frac{n}{\lambda})$, we would like $\tilde{f}(x) = \frac{1}{\lambda}\sum_n q_n g(x - \frac{n}{\lambda})$with $q_n \in \{-1, 1\}$ so $|f(t) - \tilde{f}(t)|$ is* small.

Typical Alphabets $\mathcal{A}_\delta$, i.e. set of all possible q_n. Let $\frac{1}{2^k} = \delta$ and define

$$\mathcal{A}_\delta = \{\pm\frac{1}{2^k}, \pm\frac{2}{2^k}, \pm\frac{3}{2^k} \cdots \pm 1\},$$

or $k + 1$-bit quantization. Audio encoding is typically 8, 16 or 32 bits.

Quantizer $Q_\delta(t)$ $Q_\delta(t) = \arg\min_{r\in\mathcal{A}_\delta} |t - r| = \delta\lfloor\frac{t}{\delta} + \frac{1}{2}\rfloor Q_\delta(t)$ outputs $q_n \in \mathcal{A}_\delta$ closet to t

3.1. PCM

I) Pulse Code Modulation (PCM):

For $x \in \mathbb{R}^d$ and $x = \sum_n x_n e_n$, then $q_n = Q_\delta(x_n)$

PCM White Noise Hypothesis(WNH) $x_n - Q_\delta(x_n)$ are assumed to be i.i.d in $[-1, 1)$

Commonly used in practice. More immune to chip error then binary approximation

- Under WNH: TIGHT FRAMES MINIMIZE Mean square Error (MSE): [10]‘
- WNH asymptotically true as size of alphabet increases [4,16, 17] **Works well with large alphabets!**

3.2. Sigma Delta

II) Sigma Delta($\Sigma\Delta$) scheme.

First order $\Sigma\Delta$ We introduce a state variable u_n and let $0 = u_0$ then we find inductively $u_n = x_n - q_n + u_{n-1}$ and $q_n = Q_\delta(x_n + u_{n-1})$ or $\Delta u_n = x_n - q_n$.**Higher order** $\Sigma\Delta$ For an r^{th} order scheme $\Delta^r u_n = x_n - q_n$.

In either case the error can be represented as follows using a simple summation by parts manipulation

$$\|x - \tilde{x}\|_2 = \| \sum_{n=1}^{N-r} u_n \Delta^r f_n + \sum_{j=1}^{k} u_{N-j+1} \Delta^{r-1} f_{N-j+1} \|$$

The second sum above corresponds to what we have been referring to as boundary terms. **Stable $\Sigma\Delta$ schemes** are ones for which $|u_n| < C$ as size of frame goes to ∞. All schemes considered in this note are stable ones.

4. MSE

In general higher order $\Sigma\Delta$ schemes are expected to correspond to a higher decay rate. However, in [18] it is shown that even for higher order schemes one cannot “robustly” expect more than $1/N^2$ error with the canonical dual for natural classes of frames, such as the harmonic frames. That is, in the case of harmonic frames, some oversampling rates achieve the correct approximation error but in large part it is not achieved. For this reason we look to alternate duals in the reconstruction to achieve the desired approximation rate of the algorithm. In practice, engineers make a white noise assumption for design purposes. Given the success using the WNH, we will examine the MSE problem under this setting. We treat the u_n as if they were i.i.d. This motivates the following definitions, as we will see later. In a slight abuse of notation, we will also refer to a matrix as a frame where the frame elements constitute the rows or columns of the matrix.

DEFINITION 4.1. *Given a frame $F = \{f_n\}_{n=1}^N$ for $\mathbb{R}^d$ define the **MSE Frame Variation** of F by $\sigma_2(F) = \|DF^*\|_{FR}$*
where D is the difference matrix

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 \\ & & & \ddots & \ddots & \\ 0 & 0 & \cdots & 0 & 1 & -1 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}_{N \times N}$$

and $\|A\|_{FR}$ is the Frobenius norm of a matrix A. Similarly for r^{th} order schemes we get $\sigma_2^r(F) = \|D^r F^*\|_{FR}$

DEFINITION 4.2. *For a fixed frame E we refer to the dual that minimizes $\sigma_2^r(F)$as the r^{th}* **order Sobolev Dual Frame**.

The Sobolev dual is introduced in [5] where the following theorems about it are shown.

THEOREM 4.3 (5). *For a given Frame E the Sobolev dual exits and is unique.*

THEOREM 4.4 (5). *Given a frame E, the r^{th} Sobolev dual, F, is the minimizer of $\|D^r F^*\|_{op}$, where $\|A\|_{op}$ is the operator norm of a matrix A, over all dual frames of E.*

THEOREM 4.5 (5). *Let F be the r^{th} Sobolev dual of a uniformly sampled frame path $E^*(t) = [e_1(t), e_2(t), \cdots e_d(t)]$ where $e_i(t)$ are linearly independent. If $\tilde{x} = F\mathbf{q}$ where $\mathbf{q^T} = [q_1, q_2, \cdots, q_N]$ and q_i are obtained from a stable r^{th} order $\Sigma\Delta$ scheme then $\|x - \tilde{x}\| = O(\frac{1}{N^r})$. That is the pointwise approximation error is asymptotic to $\frac{1}{N^r}$.*

Now we present the result for MSE. The equalities in the heart of the proof follow directly from the WNH, i.e.,the fact that the u_n are i.i.d.

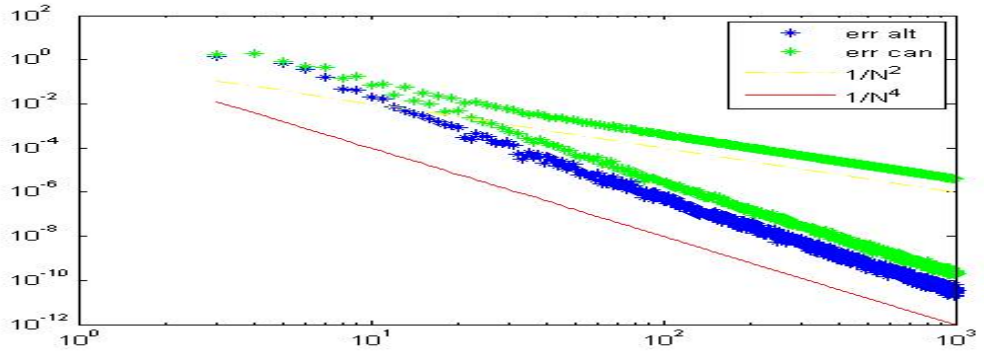
THEOREM 4.6. *Under the WNH design criteria the r^{th} Sobolev dual minimizes the MSE of a stable r^{th} order $\Sigma\Delta$ scheme.*

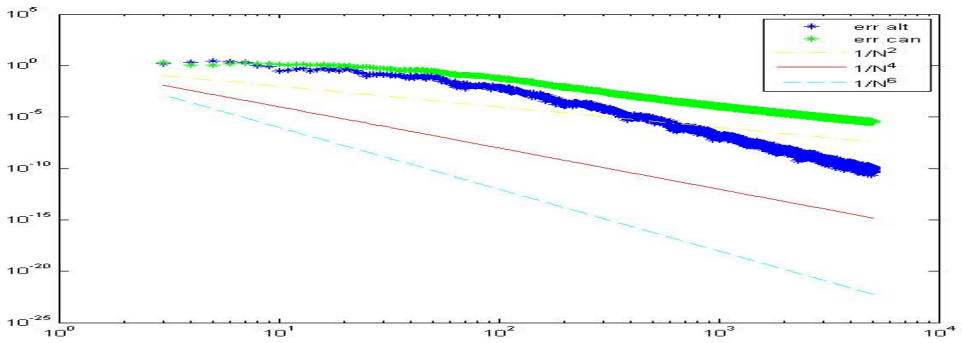
Proof. The argument for the rth case is nearly identical to r=1 so for simplicity we prove the result for $r = 1$. Let $E = \{e_n\}_{n=1}^N$ be a frame and let $\{f_n\}_{n=1}^N$ be any dual frame.

$$\begin{aligned} E\|x - \tilde{x}\|^2 =& E\left\langle \sum_{n=1}^{N-1} u_n(f_n - f_{n+1}) + u_N f_N, \sum_{m=1}^{N-1} u_m(f_m - f_{m+1}) + u_N f_N \right\rangle \\ =& \frac{\delta^2}{12}\left(\sum_{n=1}^{N-1} \|f_n - f_{n+1}\|^2 + \|f_N\|^2\right) \\ =& \frac{\delta^2}{12}\|DF^*\|_{FR} \end{aligned}$$

Here, $\frac{\delta^2}{12}$corresponds to the variance. Since δ is fixed by the choice of quantizer, Q_δ, it is clear that the Sobolev dual is the minimizer. □

Below we show show some simulations for reconstructing with the canonical dual v.s. the r^{th} Sobolev dual for both second and third order schemes where the original frame is the harmonic frame. This is a log-log plot with the x-axis corresponding to the oversampling rate and the y-axis corresponds to error. There appears to be two lines for the canonical dual in the second order scheme. This can be attributed to the fact that, for even and odd oversampling the so called boundary terms are quite different (see [2,3]).

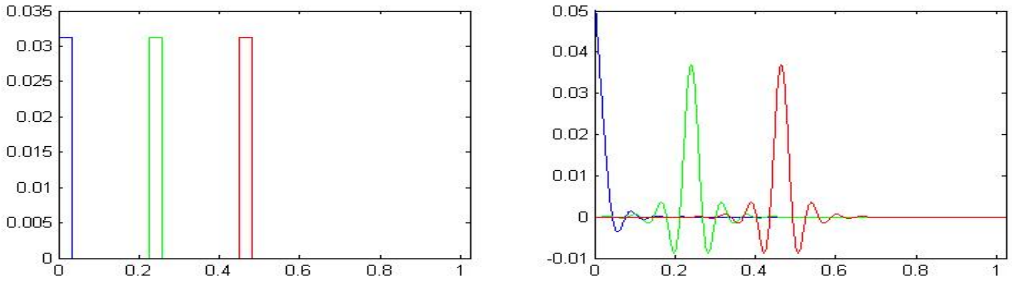




5. SOBOLEV DUALS FOR SPECIFIC FRAME PATHS

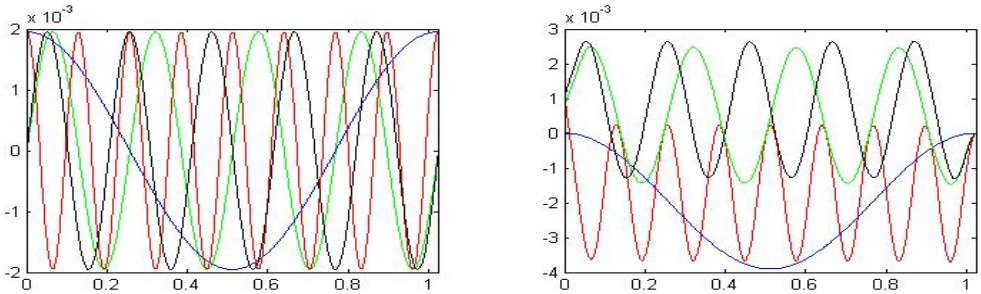
Computing the Sobolev duals for specific frames is not difficult. The authors have used two approaches with equal success. The first is to solve a Lagrange Multiplier problem, minimizing $\|D^r F^*\|$ with the constraint that F must be a dual to the frame E. The second is to use the fact that canonical duals are often minimizers and so one looks for the canonical dual of the frame ED^{-1}. This reduces to inverting the matrix $ED^{-1}D^{-1*}E^*$ and computing $(ED^{-1}D^{-1*}E^*)^{-1}ED^{-1}$. Here are a few useful examples with the sampling frame and a harmonic frame. We use the notion that these frames may be obtained from sampling continuous functions at regular intervals. Below are the paths of the original frame (up to a constant since the frames are tight) and the path of the Sobolev duals.

Component functions $\epsilon_1(t)$, $\epsilon_8(t)$, and $\epsilon_{15}(t)$ of Canonical dual (left) and $f_1(t)$, $f_8(t)$, and $f_{15}(t)$ of Alt dual(right) of sampling frame for $\mathbb{R}^{32}$



,

Component functions $\epsilon_1(t)$, $\epsilon_8(t)$, $\epsilon_1 0(t)$ and $\epsilon_{15}(t)$ of Canonical dual (left) and $f_1(t)$, $f_8(t)$, $f_1 0(t)$ and $f_{15}(t)$ of Alt dual(right) of harmonic frame for $\mathbb{R}^{32}$

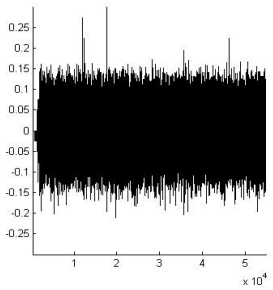
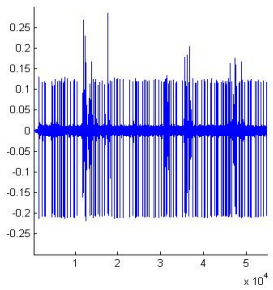


6. APPLICATION

In this section we will compare the SNR of reconstructing some audio clips with other duals versus the Sobolev dual. We keep in mind that for specific audio signals it is almost certain that one can find a dual that works better then any of the ones we use for reconstruction below, but our goal here is to find duals that work well and are not completely dependent on a single signal. First we will compare the Sobolev dual to the canonical dual. For all cases below the original frame is the sampling frame, or the so called repeated orthonormal bases. To compute SNR we are using $10\log_{10}(\frac{\sigma^2}{\sigma_e^2})$where σ^2 is the variance of the original signal and σ_e^2 is the variance of the error of the quantization. Due to limited computing power, each signal was segmented into sections of length 256 and an oversampling rate of 16 was used for the sigma delta quantization.

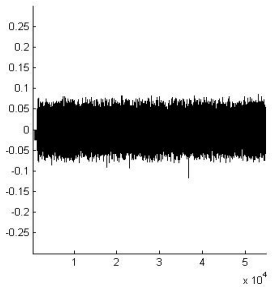
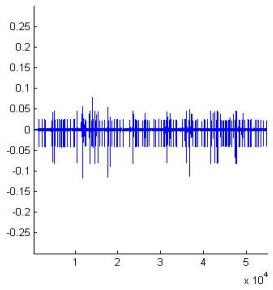
Canonical dual v.s. Sobolev dual

Quantized with same 1 bit 2^{nd} order $\Sigma\Delta$ scheme using sampling frame path with oversampling rate 16. Left: error of reconstruction with canonical dual. Right: error of reconstruction with Sobolev dual. Signal is 266 KB audio file originally encoded at 16 bit PCM.

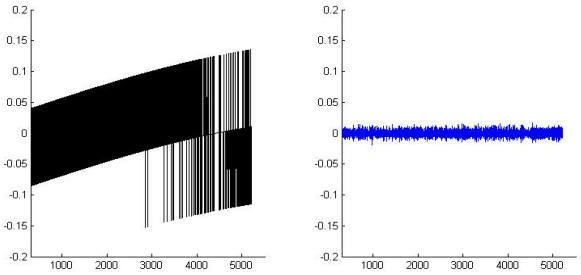
SNR with canonical dual **14.2040** SNR with Sobolev dual **26.5599**

Quantized with same 3 bit 2^{nd} order $\Sigma\Delta$ scheme using sampling frame path with oversampling rate 16. Left: error of reconstruction with canonical dual. Right: error of reconstruction with Sobolev dual. Signal is 266 KB audio file originally encoded at 16 bit PCM.

SNR with canonical dual **21.1036** SNR with Sobolev **39.4178**

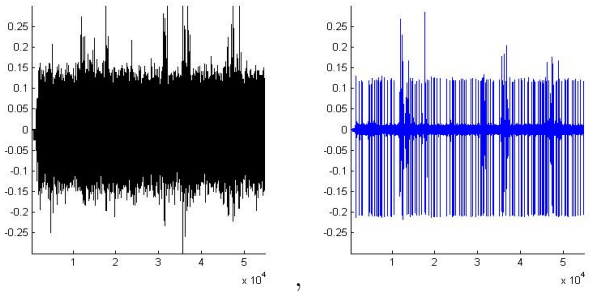
Quantized with same 1 bit 2^{nd} order $\Sigma\Delta$ scheme using sampling frame path with oversampling rate 16. Left: error of reconstruction with canonical window. Right: error of reconstruction with Sobolev dual. Signal is $y = sin(.0021x)/8 + .4$.



SNR with canonical dual**-4.9261** SNR with Sobolev dual **15.7931**

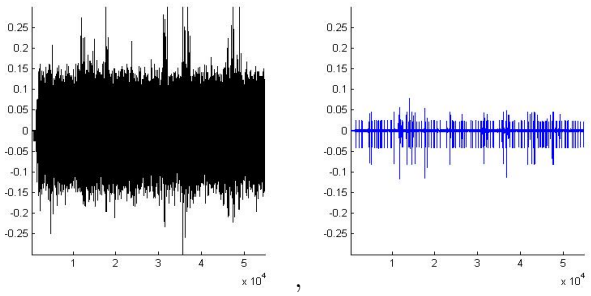
Blackman window v.s. Sobolev dual

Quantized with same 1 bit 2^{nd} order $\Sigma\Delta$ scheme using sampling frame path with oversampling rate 16. Left: error of reconstruction with Blackman. Right: error of reconstruction with Sobolev dual. Signal is 266 KB audio file originally encoded at 16 bit PCM.



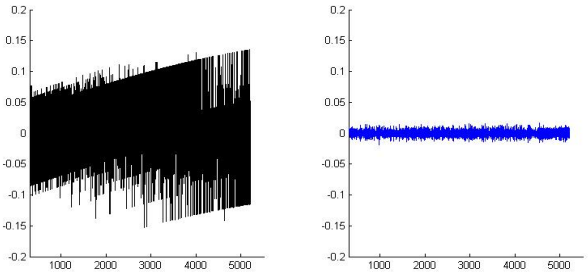
SNR with Blackman window **16.9664** SNR with Sobolev dual **26.5599**

Quantized with same 3 bit 2^{nd} order $\Sigma\Delta$ scheme using sampling frame path with oversampling rate 16. Left: error of reconstruction with Blackman window. Right: error of reconstruction with Sobolev dual. Signal is 266 KB audio file originally encoded at 16 bit PCM.



SNR with Blackman Window **32.6597** SNR with Sobolev dual **39.4178**

Quantized with same 1 bit 2^{nd} order $\Sigma\Delta$ scheme using sampling frame path with oversampling rate 16. Left: error of reconstruction with Blackman window. Right: error of reconstruction with Sobolev dual. Signal is $y = sin(.0021x)/8 + .4$.



,

SNR with Blackman Window **-3.3492**

SNR with Sobolev dual **15.7931**

REFERENCES

1. J.J. Benedetto, A.M. Powell, Ö. Yılmaz, Sigma-Delta quantization and finite frames, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 3, Montreal, QC, Canada, May 2004, pp. 937–940.
2. J.J. Benedetto, A.M. Powell, Ö. Yılmaz, Sigma-Delta ($\Sigma\Delta$) quantization and finite frames, *IEEE Transactions on Information Theory*, **52** (2006), no.5, 1990–2005.
3. J.J. Benedetto, A.M. Powell, Ö. Yılmaz, Second order Sigma-Delta ($\Sigma\Delta$) quantization of finite frame expansions, *Applied and Computational Harmonic Analysis*, **20** (2006), no.1, 126–148.
4. W. Bennett, Spectra of quantized signals, *Bell System Technical Journal*, **27** (1949), 446–472.
5. J. Blum, M.C. Lammers, A.M. Powell, Ö. Yılmaz, Sobolev Duals for frames and Sigma Delta Quantization, *Preprint*.
6. B.G. Bodmann, V.I. Paulsen, Frame paths and error bounds for sigma-delta quantization *Applied and Computational Harmonic Analysis*, **22** (2007), no.2, 176–197.
7. B.G. Bodmann, V.I. Paulsen and S.A. Abdulbaki Smooth frame path termination for higher order Sigma-Delta Quatization *to appear JFAA*
8. H. Bölcskei, F. Hlawatsch, Noise reduction in oversampled filter banks using predictive quantization, *IEEE Transactions on Information Theory*, **47** (2001), no.1, 155–172.
9. I. Daubechies, R. DeVore, Approximating a bandlimited function from very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order, *Annals of Mathematics (2)*, **158** (2003), no. 2, 679–710.
10. V. Goyal, J. Kovačević, J. Kelner, Quantized frame expansions with erasures, *Applied and Computational Harmonic Analysis*, **10** (2001), 203–233.
11. V. Goyal, M. Vetterli, N. Thao, Quantized overcomplete expansions in $\mathbb{R}^n$, *IEEE Transactions on Information Theory*, **44** (1998), no.1, 16–31.
12. C.S. Güntürk, Approximating a bandlimited function using very coarsely quantized data: improved error estimates in sigma-delta modulation, *Journal of the American Mathematical Society*, **17** (2004), no.1, 229–242.
13. Güntürk, C. Sinan, Approximating a bandlimited function using very coarsely quantized data: improved error estimates in sigma-delta modulation, *J. Amer. Math. Soc.*, **17** 2004, no. 1, 229–242.
14. Güntürk, C. Sinan and Thao, Nguyen T, Ergodic dynamics in sigma-delta quantization: tiling invariant sets and spectral analysis of error, *Adv. in Appl. Math.*, **34** 2005, no. 3,523–560.
15. R. Gray, Quantization noise spectra, *IEEE Transactions on Information Theory*, **36** (1990), no.6, 1220–1244.

16. R. Gray, Quantization Noise in $\Delta\Sigma$ A/D Converters Delta-Sigma Data Converters: Theory, Design, and Simulation. Steven R. Norsworthy (Editor), Richard Schreier (Editor), Gabor C. Temes (Editor) Wiley-IEEE Press (October 14, 1996)
17. D. Jimenez, L. Wang, Y. Wang, White noise hypothesis for unifrom quantization errors, *SIAM J. MATH ANAL.* **38** (2007) no.6, 2042-2056.
18. M.C. Lammers, A. Powell, O. Yilmaz Alternative dual frames for digital-to-analog conversion *preprint*
19. D. Marco, D. Neuhoff, The validity of the additive noise model for uniform scalar quantizers, *IEEE Transactions on Information Theory,* **51** (2005), no.5, 1739–1755.
20. A. Sripad, D. Snyder, A necessary and sufficient condition for quantization errors to be uniform and white, *IEEE Transactions on Acoustics, Speech and Signal Processing,* **25** (1977), no. 5, 442–448.
21. Ö. Yılmaz, Stability analysis for several second-order sigma-delta methods of coarse quantization of bandlimited functions, *Constructive Approximation,* **18** (2002), no.4, 599–623.
22. G. Zimmermann, Normalized tight frames in finite dimensions, in: K. Jetter, W. Haussmann, M. Reimer (Eds.), Recent Progress in Multivariate Approximation, Birkhäuser, 2001.