Quantization noise shaping on arbitrary frame expansions

Petros T. Boufounos and Alan V. Oppenheim*
Massachusetts Institute of Technology
Digital Signal Processing Group
77 Massachusetts Avenue, Rm. 36-631
Cambridge, MA 02139
{petrosb,avo}@mit.edu

July 26, 2005

Abstract

Quantization noise shaping is commonly used in oversampled A/D and D/A converters with uniform sampling. This paper considers quantization noise shaping for arbitrary finite frame expansions based on generalizing the view of first-order classical oversampled noise shaping as a compensation of the quantization error through projections. Two levels of generalization are developed, one a special case of the other, and two different cost models are proposed to evaluate the quantizer structures. Within our framework, the synthesis frame vectors are assumed given, and the computational complexity is in the initial determination of frame vector ordering, carried out off-line as part of the quantizer design. We consider the extension of the results to infinite shift-invariant frames and consider in particular filtering and oversampled filter banks.

1 Introduction

Quantization methods for frame expansions have received considerable attention in the last few years. Simple scalar quantization applied independently on each frame expansion coefficient, followed by linear reconstruction is well known to be suboptimal [8, 7]. Several algorithms have been proposed that improve performance although with significant complexity either at the quantizer [9] or

*This work was supported in part by: participation in the Advanced Sensors Collaborative Technology Alliance (CTA) sponsored by the U.S. Army Research Laboratory under Cooperative agreement DAAD19-01-2-008, the Texas Instruments Leadership University Consortium Program, BAE Systems Inc., and MIT Lincoln Laboratory. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government.

in the reconstruction method [9, 11]. More recently, frame quantization methods inspired by uniform oversampled noise shaping (referred to generically as Sigma-Delta noise shaping) have been proposed for finite uniform frames [1, 2] and for frames generated by oversampled filterbanks [3]. In [1, 2] the error due to the quantization of each expansion coefficient is subtracted from the next coefficient. The method is algorithmically similar to classical first order noise shaping and uses a quantity called frame variation to determine the optimal ordering of frame vectors such that the quantization error is reduced. In [3] higher order noise shaping is extended to oversampled filterbanks using a predictive approach. That solution performs higher order noise shaping, where the error is filtered and subtracted from the subsequent frame coefficients.

In this paper we view noise shaping as compensation of the error resulting from quantizing each frame expansion coefficient through a projection onto the space defined by another synthesis frame vector. This requires only knowledge of the synthesis frame set and a pre-specified ordering and pairing for the frame vectors. Instead of attempting a purely algorithmic generalization, we incorporate the use of projections and explore the issue of frame vector ordering. Our method improves the average quantization error even if the frame vector ordering is not optimal. However, we also demonstrate the benefits from determining the optimal ordering. The theoretical framework we present provides a design method for noise shaping quantizers under the cost functions presented. The generalization we propose improves the error in reconstruction due to quantization even for non-redundant frame expansions (i.e. a basis set) when the frame vectors are non-orthogonal. This paper elaborates and expands on [4].

In section 2 we present a brief summary of frame representations to establish notation and we describe classical first-order Sigma-Delta quantizers in the terminology of frames. In section 3 we propose two generalizations, which we refer to as the sequential quantizer and the tree quantizer, both assuming a known ordering of the frame vectors. Section 4 explores two different cost models for evaluating the quantizer structures and determining the frame vector ordering. The first is based on a stochastic representation of the error and the second on deterministic upper bounds. In section 5 we determine the optimal ordering of coefficients assuming the cost measures in section 4 and show that for Sigma-Delta noise shaping, the natural (time-sequential) ordering is optimal. We also show that for finite frames the determination of frame vector ordering can be formulated in terms of known problems in graph theory.

In section 6 we consider cases where the projection is restricted and the connection to the work in [1, 2]. Furthermore, we examine the natural extension to the case of higher order quantization. Section 7 presents experimental results on finite frames that verify and validate the theoretical ones. In section 8 we discuss infinite frame expansions. We apply the results to infinite shift invariant frames, and view filtering and classical noise shaping as an example. We also consider the case of reconstruction filterbanks, and how our work relates to [3].

2 Concepts and Background

In this section we present a brief summary of frame expansions to establish notation, and we describe oversampling in the context of frames.

2.1 Frame representation and Quantization

A vector $\mathbf{x}$ in a space $\mathcal{W}$ of finite dimension N is represented with the finite frame expansion:

$$\mathbf{x} = \sum_{k=1}^{M} a_k \mathbf{f}_k, \;\; a_k = \langle \mathbf{x}, \underline{\mathbf{f}}_k \rangle. \tag{1}$$

The space $\mathcal{W}$ is spanned by both sets: the synthesis frame vectors $\{\mathbf{f}_k, k = 1, \ldots, M\}$, and the analysis frame vectors $\{\underline{\mathbf{f}}_k, k = 1, \ldots, M\}$. This condition ensures that $M \geq N$. Details on the relationships of the analysis and synthesis vectors can be found in a variety of texts such as [8, 10]. The ratio $r = M/N$ is referred to as the *redundancy* of the frame. The equations above hold for infinite dimensional frames, with an additional constraint that ensures the sum converges for all $\mathbf{x}$ with finite length. An analysis frame is referred to as *uniform* if all the frame vectors have the same magnitude, i.e. $||\underline{\mathbf{f}}_k|| = ||\underline{\mathbf{f}}_l||$ for all k and l. Similarly, a synthesis frame is uniform if $||\mathbf{f}_k|| = ||\mathbf{f}_l||$ for all k and l.

The coefficients a_k above are scalar, continuous quantities. In order to digitally process, store, or transmit them, they need to be quantized. The simplest quantization strategy, which we call direct scalar quantization, is to quantize each one individually to $\hat{a}_k = Q(a_k) = a_k + e_k$, where $Q(\cdot)$ denotes the quantization function and e_k the quantization error for each coefficient. The total additive error vector from this strategy is equal to

$$\mathcal{E} = \sum_{k=1}^{M} e_k \mathbf{f}_k. \tag{2}$$

It is easy to show that if the frame forms an orthonormal basis, then direct scalar quantization is optimal in terms of minimizing the error magnitude. However, this is not the case for all other frame expansions [1, 2, 3, 5, 7, 8, 9, 11]. Noise shaping is one of the possible strategies to reduce the error magnitude. In order to generalize noise shaping to arbitrary frame expansions, we first present traditional oversampling and noise shaping formulated in frame terms.

2.2 Sigma-Delta Noise shaping

Oversampling in time of bandlimited signals is a well studied class of frame expansions. A signal $x[n]$ or $x(t)$ is upsampled or oversampled to produce a sequence a_k. In the terminology of frames, the upsampling operation is a frame expansion in which $\underline{\mathbf{f}}_k[n] = r\mathbf{f}_k[n] = \mathrm{sinc}(\pi(n-k)/r)$, with $\mathrm{sinc}(x) = \sin(x)/x$.

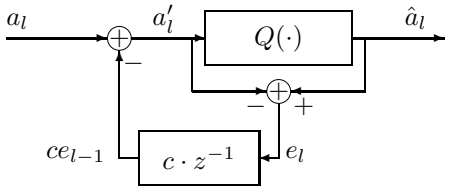

Figure 1: Traditional first order noise shaping quantizer

The sequence a_k is the corresponding ordered sequence of frame coefficients:

$$a_k = \langle x[n], \underline{\mathbf{f}}_k[n] \rangle = \sum_n x[n] \text{sinc}(\pi(n-k)/r) \tag{3}$$

$$x[n] = \sum_k a_k \mathbf{f}_k[n] = \sum_k a_k \frac{1}{r} \text{sinc}(\pi(n-k)/r). \tag{4}$$

Similarly for oversampled continuous time signals:

$$a_k = \langle x(t), \underline{\mathbf{f}}_k(t) \rangle = \int_{-\infty}^{+\infty} x(t) \frac{r}{T} \text{sinc}(\frac{\pi r t}{T} - \pi k) \tag{5}$$

$$x(t) = \sum_k a_k \mathbf{f}_k(t) = \sum_k a_k \text{sinc}(\frac{\pi r t}{T} - \pi k), \tag{6}$$

where T is the Nyquist sampling period for $x(t)$.

Sigma-Delta quantizers can be represented in a number of equivalent forms [5]. The representation shown in figure 1 most directly represents the view that we extend to general frame expansions. Performance of Sigma-Delta quantizers is sometimes analyzed using an additive white noise model for the quantization error [5]. Based on this model it is straightforward to show that the in-band quantization noise power is minimized when the scaling coefficient c is chosen to be $c = \text{sinc}(\pi/r)$[1].

We view the process in figure 1 as an iterative process of coefficient quantization followed by error projection. The quantizer in the figure quantizes a'_l to $\hat{a}_l = a'_l + e_l$. Consider $x_l[n]$, such that the coefficients up to a_{l-1} have been quantized and e_{l-1} has already been scaled by c and subtracted from a_l to produce a'_l:

$$x_l[n] = \sum_{k=-\infty}^{l-1} \hat{a}_k \mathbf{f}_k[n] + a'_l \mathbf{f}_l[n] + \sum_{k=l+1}^{+\infty} a_k \mathbf{f}_k[n] \tag{7}$$

$$= x_{l+1}[n] + e_l(\mathbf{f}_l[n] - c \cdot \mathbf{f}_{l+1}[n]). \tag{8}$$

The incremental error $e_l(\mathbf{f}_l[n] - c \cdot \mathbf{f}_{l+1}[n])$ at the l^{th} iteration of (8) is minimized if we pick c such that $c \cdot \mathbf{f}_{l+1}[n]$ is the projection of $\mathbf{f}_l[n]$ onto $\mathbf{f}_{l+1}[n]$:

$$c = \langle \mathbf{f}_l[n], \mathbf{f}_{l+1}[n] \rangle / ||\mathbf{f}_{l+1}[n]||^2 = \text{sinc}(\pi/r). \tag{9}$$

[1] With typical oversampling ratios, this coefficient is close to unity and is often chosen as unity for computational convenience.

This choice of c projects to $\mathbf{f}_{l+1}[n]$ the error due to quantizing a_l and compensates for this error by modifying a_{l+1}. Note that the optimal choice of c in (9) is the same as the optimal choice of c under the additive white noise model for quantization.

Minimizing the incremental error is not necessarily optimal in terms of minimizing the overall quantization error. It is, however, optimal in terms of the two cost functions which we describe in section 4. Before we examine these cost functions we generalize first order noise shaping to general frame expansions.

3 Noise shaping on Frames

In this section we propose two generalizations of the discussion of section 2.2 to arbitrary finite-frame representations of length M. Throughout the discussion in this section we assume the ordering of the synthesis frame vectors $(\mathbf{f}_1, \ldots, \mathbf{f}_M)$, and correspondingly the ordering of the synthesis coefficients $(a_1, \ldots, a_M)$ has already been determined.

We examine the ordering of the frame vectors in section 5. However, we should emphasize that the execution of the algorithm and the ordering of the frame vectors are distinct issues. The optimal ordering can be determined once, off-line, in the design phase. The ordering only depends on the properties of the synthesis frame, not the data or the analysis frame.

3.1 Single coefficient quantization

To illustrate our approach, we consider quantizing the first coefficient a_1 to $\hat{a}_1 = a_1 + e_1$, with e_1 denoting the additive quantization error. Equation (1) then becomes:

$$\mathbf{x} = \hat{a}_1\mathbf{f}_1 + \sum_{k=2}^{M} a_k\mathbf{f}_k - e_1\mathbf{f}_1 \tag{10}$$

$$= \hat{a}_1\mathbf{f}_1 + a_2\mathbf{f}_2 + \sum_{k=3}^{M} a_k\mathbf{f}_k - e_1c_{1,2}\mathbf{f}_2 - e_1(\mathbf{f}_1 - c_{1,2}\mathbf{f}_2). \tag{11}$$

As in (8), the norm of $e_1(\mathbf{f}_1 - c_{1,2}\mathbf{f}_2)$ is minimized if $c_{1,2}\mathbf{f}_2$ is the projection of $\mathbf{f}_1$ onto $\mathbf{f}_2$:

$$c_{1,2}\mathbf{f}_2 = \langle \mathbf{f}_1, \mathbf{u}_2\rangle \mathbf{u}_2 \tag{12}$$

$$= \langle \mathbf{f}_1, \frac{\mathbf{f}_2}{||\mathbf{f}_2||}\rangle \frac{\mathbf{f}_2}{||\mathbf{f}_2||} \tag{13}$$

$$\Rightarrow c_{1,2} = \frac{\langle \mathbf{f}_1, \mathbf{u}_2\rangle}{||\mathbf{f}_2||} = \frac{\langle \mathbf{f}_1, \mathbf{f}_2\rangle}{||\mathbf{f}_2||^2}, \tag{14}$$

where $\mathbf{u}_k = \mathbf{f}_k/||\mathbf{f}_k||$ are unit vectors in the direction of the synthesis vectors. Next, we incorporate the term $-e_1c_{1,2}\mathbf{f}_2$ in the expansion by updating a_2:

$$a_2' = a_2 - e_1c_{1,2}. \tag{15}$$

After the projection, the residual error is equal to $e_1(\mathbf{f}_1 - c_{1,2}\mathbf{f}_2)$. To simplify this expression, we define $\mathbf{r}_{1,2}$ to be the direction of the residual error, and $e_1\tilde{c}_{1,2}$ to be the error amplitude:

$$\mathbf{r}_{1,2} = (\mathbf{f}_1 - c_{1,2}\mathbf{f}_2)/||\mathbf{f}_1 - c_{1,2}\mathbf{f}_2|| \quad (16)$$

$$\tilde{c}_{1,2} = ||\mathbf{f}_1 - c_{1,2}\mathbf{f}_2|| = \langle \mathbf{f}_1, \mathbf{r}_{1,2} \rangle. \quad (17)$$

Thus, the residual error is $e_1\langle \mathbf{f}_1, \mathbf{r}_{1,2} \rangle \mathbf{r}_{1,2} = e_1\tilde{c}_{1,2}\mathbf{r}_{1,2}$. We refer to $\tilde{c}_{1,2}$ as the *error coefficient* for this pair of vectors.

Substituting the above, equation (11) becomes

$$\mathbf{x} = \hat{a}_1\mathbf{f}_1 + a_2'\mathbf{f}_2 + \sum_{k=3}^{M} a_k\mathbf{f}_k - e_1\tilde{c}_{1,2}\mathbf{r}_{1,2}. \quad (18)$$

Equation (18) can be viewed as decomposing $e_1\mathbf{f}_1$ into the direct sum $(e_1c_{1,2}\mathbf{f}_2) \oplus (e_1\tilde{c}_{1,2}\mathbf{r}_{1,2})$ and compensating only for the first term of this sum. The component $e_1\tilde{c}_{1,2}\mathbf{r}_{1,2}$ is the final quantization error after one step is completed.

Note that for any pair of frame vectors the corresponding error coefficient $\tilde{c}_{k,l}$ is always positive. Also, if we assume a uniform synthesis frame, there is a symmetry in the terms we defined, i.e. $c_{k,l} = c_{l,k}$ and $\tilde{c}_{k,l} = \tilde{c}_{l,k}$, for any pair $k \neq l$.

3.2 Sequential Noise Shaping Quantizer

The process in section 3.1 is iterated by quantizing the next (updated) coefficient until all the coefficients have been quantized. Specifically, the algorithm continues as follows:

1. Quantize coefficient k by setting $\hat{a}_k = Q(a_k')$.
2. Compute the error $e_k = \hat{a}_k - a_k'$.
3. Update the next coefficient a_{k+1} to $a_{k+1}' = a_{k+1} - e_k c_{k,k+1}$, where

$$c_{k,l} = \frac{\langle \mathbf{f}_k, \mathbf{f}_l \rangle}{||\mathbf{f}_l||^2}. \quad (19)$$

4. Increase k and iterate from step 1 until all the coefficients have been quantized.

We refer to this procedure as the *sequential* first order noise shaping quantizer.

Every iteration of the sequential quantization contributes $e_k\tilde{c}_{k,k+1}\mathbf{r}_{k,k+1}$ to the total quantization error, where

$$\mathbf{r}_{k,l} = \frac{\mathbf{f}_k - c_{k,l}\mathbf{f}_l}{||\mathbf{f}_k - c_{k,l}\mathbf{f}_l||}, \text{ and} \quad (20)$$

$$\tilde{c}_{k,l} = ||\mathbf{f}_k - c_{k,l}\mathbf{f}_l||. \quad (21)$$

Since the frame expansion is finite, we cannot compensate for the quantization error of the last step $e_M \mathbf{f}_M$. Thus, the total error vector is

$$\mathcal{E} = \sum_{k=1}^{M-1} e_k \tilde{c}_{k,k+1} \mathbf{r}_{k,k+1} + e_M \mathbf{f}_M. \tag{22}$$

Note that $\tilde{c}_{k,l}\mathbf{r}_{k,l}$ is the residual from the projection of $\mathbf{f}_k$ onto $\mathbf{f}_l$, and therefore it has magnitude less than or equal to $\mathbf{f}_k$. Specifically, for all k and l,

$$\tilde{c}_{k,l} \leq ||\mathbf{f}_k||, \tag{23}$$

with equality holding if and only if $\mathbf{f}_k$ is orthogonal to $\mathbf{f}_l$. Furthermore note that since $\tilde{c}_{k,l}$ is the magnitude of a vector it is always nonnegative.

3.3 The Tree Noise Shaping Quantizer

The sequential quantizer can be generalized by relaxing the sequence of error assignments: Again, we assume that the coefficients have been pre-ordered and that the ordering defines the sequence in which coefficients are quantized. In this generalization, we associate with each ordered frame vector $\mathbf{f}_k$ another, not necessarily adjacent, frame vector $\mathbf{f}_{l_k}$ further in the sequence (and, therefore, for which the corresponding coefficient has not yet been quantized) to which the error is projected using (15). With this more general approach some frame vectors can be used to compensate for more than one quantized coefficient.

In terms of the algorithm presented in section 3.2, step 3 changes to:

3. Update a_{l_k} to $a'_{l_k} = a_{l_k} - e_k c_{k,l_k}$, where $c_{k,l} = \frac{\langle \mathbf{f}_k, \mathbf{f}_l \rangle}{||\mathbf{f}_l||^2}$, and $l_k > k$.

The constraint $l_k > k$ ensures that a_{l_k} is further in the sequence than a_k. For finite frames, this defines a tree, in which every node is a frame vector or associated coefficient. If a coefficient a_k uses coefficient a_{l_k} to compensate for the error, then a_k is a direct child of a_{l_k} in that tree. The root of the tree is the last coefficient to be quantized, a_M.

We refer to this as the *tree* noise shaping quantizer. The sequential quantizer is, of course, a special case of the tree quantizer where $l_k = k + 1$.

The resulting expression for $\mathbf{x}$ is given by:

$$\mathbf{x} = \sum_{k=1}^{M} \hat{a}_k \mathbf{f}_k - \sum_{k=1}^{M-1} e_k \tilde{c}_{k,l_k} \mathbf{r}_{k,l_k} - e_M \mathbf{f}_M \tag{24}$$

$$= \hat{\mathbf{x}} - \sum_{k=1}^{M-1} e_k \tilde{c}_{k,l_k} \mathbf{r}_{k,l_k} - e_M ||\mathbf{f}_M|| \mathbf{u}_M, \tag{25}$$

where $\hat{\mathbf{x}}$ is the quantized version of $\mathbf{x}$ after noise shaping, and the e_k are the quantization errors in the coefficients *after* the corrections from the previous iterations have been applied to a_k. Thus, the total error of the process is:

$$\mathcal{E} = \sum_{k=1}^{M-1} e_k \tilde{c}_{k,l_k} \mathbf{r}_{k,l_k} + e_M \mathbf{f}_M. \tag{26}$$

4 Error Models and Analysis

In order to compare and design quantizers, we need to be able to compare the magnitude of the error in each. However, the error terms e_k in equations (2), (22), and (26) are data dependent in a very non-linear way. Furthermore, due to the error projection and propagation performed in noise shaping, the coefficients being quantized at every step are different for the different quantization strategy. Therefore, for each k, e_k is different among the equations (2), (22), and (26), making the precise analysis and comparison even harder. In order to compare quantizer designs we need to evaluate them using cost functions that are independent of the data.

To simplify the problem further, we focus on cost measures for which the incremental cost at each step is independent of the whole path and the data. We refer to these as *incremental* cost functions. In this section we examine two such models, one stochastic and one deterministic. The first cost function is based on the white noise model for quantization, while the second provides a guaranteed upper bound for the error. Note that for the rest of this development we assume linear quantization, with Δ denoting the interval spacing of the linear quantizer. We also assume that the quantizer is properly scaled to avoid overflow.

4.1 Additive Noise Model

The first cost function assumes the additive uniform white noise model for quantization error, to determine the expected energy of the error $E\{||\mathcal{E}||^2\}$. An additive noise model has previously been applied to other frame expansions [3, 9]. Its assumptions are often inaccurate, and it only attempts to describe average behavior, with no guarantees on performance comparisons or improvements for individual realizations. However it can often lead to important insights on the behavior of the quantizer.

In this model all the error coefficients e_k are assumed white and identically distributed,with variance $\Delta^2/12$, where Δ is the interval spacing of the quantizer. They are also assumed to be uncorrelated with the quantized coefficients. Thus, all error components contribute additively to the error power, resulting in:

$$E\{||\mathcal{E}||^2\} = \frac{\Delta^2}{12}\left(\sum_{k=1}^{M}||\mathbf{f}_k||^2\right), \tag{27}$$

$$E\{||\mathcal{E}||^2\} = \frac{\Delta^2}{12}\left(\sum_{k=1}^{M-1}\tilde{c}_{k,k+1}^2 + ||\mathbf{f}_M||^2\right), \text{ and} \tag{28}$$

$$E\{||\mathcal{E}||^2\} = \frac{\Delta^2}{12}\left(\sum_{k=1}^{M-1}\tilde{c}_{k,l_k}^2 + ||\mathbf{f}_M||^2\right), \tag{29}$$

for the direct, the sequential and the tree quantizer respectively.

4.2 Error magnitude upper bound

As an alternative to the cost function in section 4.1, we also consider an upper bound for the error magnitude. For any set of vectors $\mathbf{u}_i$, $||\sum_k \mathbf{u}_k|| \leq \sum_k ||\mathbf{u}_k||$, with equality only if all vectors are collinear, in the same direction. This leads to the following upper bound on the error:

$$||\mathcal{E}|| \leq \frac{\Delta}{2}\left(\sum_{k=1}^{M}||\mathbf{f}_k||\right), \tag{30}$$

$$||\mathcal{E}|| \leq \frac{\Delta}{2}\left(\sum_{k=1}^{M-1}\tilde{c}_{k,k+1}+||\mathbf{f}_M||\right), \text{ and} \tag{31}$$

$$||\mathcal{E}|| \leq \frac{\Delta}{2}\left(\sum_{k=1}^{M-1}\tilde{c}_{k,l_k}+||\mathbf{f}_M||\right), \tag{32}$$

for direct, sequential and tree quantization, respectively.

The vector $\mathbf{r}_{M-1,l_{M-1}}$ is by construction orthogonal to $\mathbf{f}_M$ and the $\mathbf{r}_{k,l_k}$ are never collinear, making the bound very loose. Thus, a noise shaping quantizer can be expected in general to perform better than what the bound suggests. Still, for the purposes of this discussion we treat this upper bound as a cost function and we design the quantizer such that this cost function is minimized.

4.3 Analysis of the Error Models

To compare the average performance of direct coefficient quantization to the proposed noise shaping we only need to compare the magnitude of the right hand side of equations (27) thru (29), and (30) thru (32) above. The cost of direct coefficient quantization computed using equations (27) and (30) does not change, even if the order in which the coefficients are quantized changes. Therefore, we can assume the ordering of the synthesis frame vectors and the associated coefficients is given, and compare the three strategies. In this section we show that for any frame vector ordering, the proposed noise shaping strategies reduce both the average error power, and the worst case error magnitude, as described using the proposed functions, compared to direct scalar quantization.

When comparing the cost functions using inequalities, the multiplicative terms $\frac{\Delta^2}{12}$ and $\frac{\Delta}{2}$, common in all equations, are eliminated, because they do not affect the monotonicity. Similarly for the final additive term $||\mathbf{f}_M||^2$ and $||\mathbf{f}_M||$, which also exists in all equations and does not affect the monotonicity of the comparison. To summarize, we need to compare the following quantities:

$$\sum_{k=1}^{M-1}||\mathbf{f}_k||^2, \quad \sum_{k=1}^{M-1}\tilde{c}_{k,k+1}^2, \text{ and } \sum_{k=1}^{M-1}\tilde{c}_{k,l_k}^2, \tag{33}$$

in terms of the average error power, and

$$\sum_{k=1}^{M-1}||\mathbf{f}_k||, \quad \sum_{k=1}^{M-1}\tilde{c}_{k,k+1}, \text{ and } \sum_{k=1}^{M-1}\tilde{c}_{k,l_k}, \tag{34}$$

in terms of the guaranteed worst case performance. These correspond to direct coefficient quantization, sequential noise shaping, and tree noise shaping respectively.

Using (23) it is easy to show that both noise shaping methods have lower cost than direct coefficient quantization for any frame vector ordering. Furthermore, we can always pick $l_k = k + 1$, and, therefore, the tree noise shaping quantizer can always achieve the cost of the sequential quantizer. Therefore, we can always find l_k such that the comparison above becomes:

$$\sum_{k=1}^{M-1} ||\mathbf{f}_k||^2 \geq \sum_{k=1}^{M-1} \tilde{c}_{k,k+1}^2 \geq \sum_{k=1}^{M-1} \tilde{c}_{k,l_k}^2, \text{ and} \tag{35}$$

$$\sum_{k=1}^{M-1} ||\mathbf{f}_k|| \geq \sum_{k=1}^{M-1} \tilde{c}_{k,k+1} \geq \sum_{k=1}^{M-1} \tilde{c}_{k,l_k}. \tag{36}$$

The relationships above hold with equality if and only if *all* the pairs $(\mathbf{f}_k, \mathbf{f}_{k+1})$ and $(\mathbf{f}_k, \mathbf{f}_{l_k})$ are orthogonal. Otherwise the comparison with direct coefficient quantization results in a strict inequality. In other words, noise shaping improves the quantization cost compared to direct coefficient quantization even if the frame is not redundant, as long as the frame is not an orthogonal basis[2]. Note that the coefficients $c_{k,l}$ are 0 if the frame is an orthogonal basis. Therefore, the feedback terms $e_k c_{k,l_k}$ in step 3 of the algorithms described in section 3 are equal to 0. In this case, the strategies in section 3 reduce to direct coefficient quantization, which can be shown to be the optimal scalar quantization strategy for orthogonal basis expansions.

We can also determine a lower bound for the cost, independent of the frame vector ordering, by picking $j_k = \text{argmin}_{l_k \neq k} \tilde{c}_{k,l_k}$. This does not necessarily satisfy the constrain $j_k > k$ of section 3.3, therefore the lower bound cannot always be met. However, if a quantizer can meet it, it is the minimum cost first order noise shaping quantizer, independent of the frame vector ordering, for both cost functions.

The inequalities presented in this section are summarized below.

For given frame ordering, $j_k = \text{argmin}_{l_k \neq k} \tilde{c}_{k,l_k}$ and some $\{l_k > k\}$:

$$\sum_{k=1}^{M} \tilde{c}_{k,j_k} \leq \sum_{k=1}^{M-1} \tilde{c}_{k,l_k} + ||\mathbf{f}_M|| \leq \sum_{k=1}^{M-1} \tilde{c}_{k,k+1} + ||\mathbf{f}_M|| \leq \sum_{k=1}^{M} ||\mathbf{f}_k||, \tag{37}$$

and

$$\sum_{k=1}^{M} \tilde{c}_{k,j_k}^2 \leq \sum_{k=1}^{M-1} \tilde{c}_{k,l_k}^2 + ||\mathbf{f}_M||^2 \leq \sum_{k=1}^{M-1} \tilde{c}_{k,k+1}^2 + ||\mathbf{f}_M||^2 \leq \sum_{k=1}^{M} ||\mathbf{f}_k||^2, \tag{38}$$

[2] An oblique basis can reduce the quantization error compared to an orthogonal one if noise shaping is used, assuming the quantizer uses the same Δ. However, more quantization levels might be necessary to ensure that the quantizer does not overflow if an oblique basis is used.

where the lower and upper bounds are independent of the frame vector ordering.

In the discussion above we showed that the proposed noise shaping reduces the average and the upper bound of the quantization error for all frame expansions. The strategies above degenerate to direct coefficient quantization if the frame is an orthogonal basis. These results hold without any assumptions on the frame, or the ordering of the frame vectors and the corresponding coefficients. Finally, we derived a lower bound for the cost of a first order noise shaping quantizer. In the next section we examine how to determine the optimal ordering and pairing of the frame vectors.

5 First Order Quantizer Design

As indicated earlier, an essential issue in first order quantizer design based on the strategies outlined in this paper is determining the ordering of the frame vectors. The optimal ordering depends on the specific set of synthesis frame vectors, but not on the specific signal. Consequently, the quantizer design (i.e. the frame vector ordering) is carried out off-line and the quantizer implementation is a sequence of projections based on the ordering chosen for either the sequential or tree quantizer.

5.1 Simple Design Strategies

An obvious design strategy is to determine an ordering and pairing of the coefficients such that the quantization of every coefficient a_k is compensated as much as possible by the coefficient a_{l_k}. This can be achieved by setting $l_k = j_k$, with $j_k = \text{argmin}_{l_k \neq k} \tilde{c}_{k,l_k}$, as defined for the lower bounds of equations (37) and (38). When this strategy is possible to implement, i.e. $j_k > k$, it results in the optimal ordering and pairing under both cost models we discussed, since it meets the lower bound for the quantization cost.

This corresponds to how a traditional Sigma-Delta quantizer works. When an expansion coefficient is quantized, the coefficients that can compensate for most of the error are the ones most adjacent. This implies that the time sequential ordering of the oversampling frame vectors is the optimal ordering for first order noise shaping (another optimal ordering is the time-reversed, i.e. the anticausal version). We examine this further in section 8.1.

Unfortunately, for certain frames, this optimal pairing might not be feasible. Still, it suggests a heuristic for a good coefficient pairing: at every step k, the error from quantizing coefficient a_k is compensated using the coefficient a_{l_k} that can compensate for most of the error, picking from all the frame vectors whose corresponding coefficients have not yet been quantized. This is achieved by setting $l_k = \text{argmin}_{l>k} \tilde{c}_{k,l}$. This, in general is not an optimal strategy, but an implementable heuristic. Optimal designs are slightly more involved and we discuss these next.

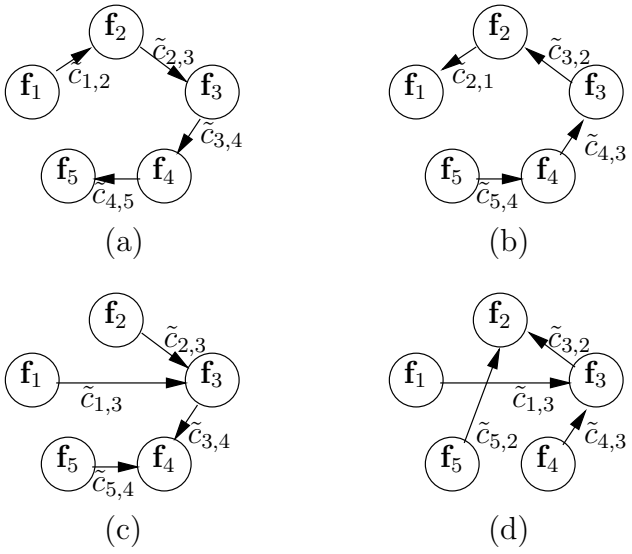

Figure 2: Examples of graph representations of first order noise shaping quantizers on a frame with five frame vectors. Note that the weights shown represent the upper bound of the quantization error. To represent the average error power the weights should be squared.

5.2 Quantization graphs and optimal quantizers

From section 3.3 it is clear that a tree quantizer can be represented as a graph—specifically, a tree—in which all the nodes of the graph are coefficients to be quantized. Similarly for a sequential quantizer, which is a special case of the tree quantizer, the graph is a linear path passing through all the nodes a_k in the correct sequence. In both cases, the graphs have edges (k, l_k), pairing coefficient a_k to coefficient a_{l_k} if and only if the quantization of coefficient a_k assigns the error to the coefficient a_{l_k}.

Figure 2 shows four examples of graph representations of first order noise shaping quantizers on a frame with five frame vectors. The top two figures, (a) and (b), demonstrate two sequential quantizers ordering the frame vectors in their natural and their reverse order respectively. In addition, parts (c) and (d) of the figure demonstrate two general tree quantizers for the same frame.

In the figure a weight is assigned to each edge. The cost of each quantizer is proportional to the total weight of the graph with the addition of the cost of the final term. For a uniform frame the magnitude of the final term is the same, independent of which coefficient is quantized last. Therefore it is eliminated when comparing the cost of quantizer designs on the same frame. Thus, designing the optimal quantizer corresponds to determining the graph with the minimum weight.

We define a graph that has the frame vectors as nodes $V = \{\mathbf{f}_1, \ldots, \mathbf{f}_M\}$ and

the edges have weight $w(k,l) = \tilde{c}_{k,l}^2$ or $w(k,l) = \tilde{c}_{k,l}$ if we want to minimize the expected error power or the upper bound of the error magnitude respectively. We call this graph the *quantization error assignment* graph. On this graph, any acyclical path that visits all the nodes—also known as a hamiltonian path—defines a first order sequential quantizer. Similarly, any tree that visits all the nodes—also known as a spanning tree—defines a tree quantizer.

The minimum cost hamiltonian path defines the optimal sequential quantizer. This can be determined by solving the *traveling salesman problem (TSP)*. The TSP is of course NP-complete in general, but has been extensively studied in the literature [6]. Similarly, the optimal tree quantizer is defined by the solution of the *minimum spanning tree* problem. This is also a well studied problem, solvable in polynomial time [6]. Since any path is also a tree, if the minimum spanning tree is a hamiltonian path, then it is also the solution to the traveling salesman problem. The results are easy to extend to non-uniform frames.

We should note that in general the optimal ordering and pairing depend on which of the two cost functions we choose to optimize for. Furthermore, we should reemphasize that this optimization is performed once, off-line, at the design stage of the quantizer. Therefore, the computational cost of solving these problems does not affect the complexity of the resulting quantizer.

6 Further Generalizations

In this section we consider two further generalizations. In section 6.1 we examine the case for which the product term is restricted. In section 6.2 we consider the case of noise shaping using more than one vector for compensation. Although a combination of the two is possible, we do not consider it in this paper.

6.1 Projection Restrictions

The development in this paper uses the product $e_k c_{k,l_k}$ to compensate for the error in quantizing coefficient a_k using coefficient a_{l_k}. Implementation restrictions often do not allow for this product to be computed to a satisfactory precision. For example, typical Sigma-Delta converters eliminate this product altogether by setting $c = 1$. In such cases, the analysis using projections breaks down. Still, the intuition and approach remains applicable.

The restriction we consider is one on the product: the coefficients c_{k,l_k} are restricted to be in a discrete set $\mathcal{A} = \{\alpha_1, ..., \alpha_K\}$. Requiring the coefficient to be an integer power of 2 or to be only ± 1 are examples of such constraints. In this case we use again the algorithms of section 3, with $c_{k,l}$ now chosen to be the coefficient in $\mathcal{A}$ closest to achieving a projection, i.e. with $c_{k,l}$ specified as:

$$c_{k,l} = \text{argmin}_{c \in \mathcal{A}} ||\mathbf{f}_k - c\mathbf{f}_l|| \tag{39}$$

As in the unrestricted case, the residual error is $e_k(\mathbf{f}_k - c_{k,l}\mathbf{f}_l) = e_k \tilde{c}_{k,l} \mathbf{r}_{k,l}$ with $\mathbf{r}_{k,l}$ and $\tilde{c}_{k,l}$ defined as in equations (20) and (21), respectively.

To apply either of the error models in section 4 we use the new $\tilde{c}_{l,l_k}$, as computed above. However, in this case, certain coefficient orderings and pairings might increase the overall error. A pairing of $\mathbf{f}_k$ with $\mathbf{f}_{l_k}$ improves the cost if and only if

$$||\mathbf{f}_k - c_{k,l_k}\mathbf{f}_{l_k}|| \leq ||\mathbf{f}_k|| \Leftrightarrow \tilde{c}_{k,l_k} \leq ||\mathbf{f}_k||, \tag{40}$$

which is no longer guaranteed to hold. Thus, the strategies described in section 5.1 need a minor modification: we only allow the compensation to take place if (40) holds. Similarly, in terms of the graphical model of section 5.2, we only allow an edge in the graph if (40) holds. Still, the optimal sequential quantizer is the solution to the TSP problem, and the optimal tree quantizer is the solution to the minimum spanning tree problem on that graph—which might now have missing edges.

The main implication of missing edges is that, depending on the frame we operate on, the graph might have disconnected components. In this case we should solve the traveling salesman problem or the minimum spanning tree on every component. Also, it is possible that, although we are operating on an oversampled frame, noise shaping is not beneficial due to the constraints. The simplest way to fix this is to always allow the choice $c_{k,l_k} = 0$ in the set $\mathcal{A}$. This ensures that (40) is always met, and therefore the graph stays connected. Thus, whenever noise shaping is not beneficial, the algorithms will pick $c_{k,l_k} = 0$ as the compensation coefficient, which is equivalent to no noise shaping. We should note that the choice of the set $\mathcal{A}$ matters. The denser the set is, the better the approximation of the projection. Thus the resulting error is smaller.

An interesting special case corresponds to removing the multiplication from the feedback loop by setting $\mathcal{A} = \{1\}$. As we mentioned before, this is a common design choice in traditional Sigma-Delta converters. Furthermore, it is the case examined in [1, 2], in which the issue of the optimal permutation is addressed in terms of the frame variation. The frame variation is defined in [1] motivated by the triangle inequality, as is the upper bound model of section 4.2. In that work it is also shown that incorrect frame vector ordering might increase the overall error, compared to direct coefficient quantization.

In this case the compensation is improving the cost if and only if $||\mathbf{f}_k - \mathbf{f}_{l_k}|| < ||\mathbf{f}_k||$. The rest of the development remains the same: we need to solve the traveling salesman problem or the minimum spanning tree problem on a possibly disconnected graph. In the example we present in section 7, the natural frame ordering becomes optimal using our cost models, yielding the same results as the frame variation criterion suggested in [1, 2]. In section 8.1 we show that when applied to classical first order noise shaping this restriction does not affect the optimal frame ordering and does not impact significantly the error power.

6.2 Higher Order quantization

Classical Sigma-Delta noise shaping is commonly done in multiple stages to achieve higher-order noise shaping. Similarly noise shaping on arbitrary frame expansions can be generalized to higher order. Unfortunately, in this case deter-

mining the optimal ordering is not as straightforward, and we do not attempt the full development in this paper. However, we develop the quantization strategy and the error modeling for a given ordering of the coefficients.

The goal of higher order noise shaping is to compensate for quantization of each coefficient using more than one coefficients. There are several possible implementations of a traditional higher order Sigma-Delta quantizers. All have a common property; the quantization error is in effect modified by a p^{th} order filter, typically with a transfer function of the form:

$$H_e(z) = (1 - z^{-1})^p \tag{41}$$

and equivalently an impulse response:

$$h_e[n] = \delta[n] - \sum_{i=1}^{p} c_i \delta[n-i]. \tag{42}$$

Thus, every error coefficient e_k additively contributes a term of the form $e_k(\mathbf{f}_k - \sum_{i=1}^{p} c_i \mathbf{f}_{k+i})$ to the output error. In order to minimize the magnitude of this contribution we need to choose the c_i such that $\sum_{i=1}^{p} c_i \mathbf{f}_{k+i}$ is the projection of $\mathbf{f}_k$ to the space spanned by $\{\mathbf{f}_{k+1}, \ldots, \mathbf{f}_{k+p}\}$. Using (41) as the system function is often preferred for implementation simplicity but it is not the optimal choice. This design choice is similar to eliminating the product in figure 1. As with first order noise shaping, it is straightforward to generalize this to arbitrary frames.

Given a frame vector ordering, we consider the quantization of coefficient a_k to $\hat{a}_k = a_k + e_k$. This error is to be compensated using coefficients a_{l_1} to a_{l_p}, with all the $l_i > k$. Thus, we project the vector $-e_k \mathbf{f}_k$ to the space $\mathcal{S}_k$, defined by the vectors $\mathbf{f}_{l_1}, \ldots, \mathbf{f}_{l_p}$. The essential part of this development is to determine a set of coefficients that multiply the error e_k in order to project it to the appropriate space.

To perform this projection we view the set $\{\mathbf{f}_l | l \in S_k\}$ as the reconstruction frame for $\mathcal{S}_k$, where $S_k = \{l_1, \ldots, l_p\}$ is the set of the indices of all the vectors that we use for compensation of coefficient a_k. Ensuring that for all $j \geq k,\ k \notin S_j$ guarantees that once a coefficient is quantized, it is not modified again.

Extending the first order quantizer notation, we denote the coefficients that perform the projection by c_{k,l,S_k}. It is straightforward to show that these coefficients perform a projection if and only if they satisfy the following equation:

$$\begin{bmatrix} \langle \mathbf{f}_{l_1}, \mathbf{f}_{l_1} \rangle & \langle \mathbf{f}_{l_1}, \mathbf{f}_{l_2} \rangle & \cdots & \langle \mathbf{f}_{l_1}, \mathbf{f}_{l_p} \rangle \\ \langle \mathbf{f}_{l_2}, \mathbf{f}_{l_1} \rangle & \langle \mathbf{f}_{l_2}, \mathbf{f}_{l_p} \rangle & \cdots & \langle \mathbf{f}_{l_1}, \mathbf{f}_{l_p} \rangle \\ \vdots & & \ddots & \vdots \\ \langle \mathbf{f}_{l_p}, \mathbf{f}_{l_1} \rangle & \langle \mathbf{f}_{l_p}, \mathbf{f}_{l_2} \rangle & \cdots & \langle \mathbf{f}_{l_p}, \mathbf{f}_{l_p} \rangle \end{bmatrix} \begin{bmatrix} c_{k,l_1,S_k} \\ c_{k,l_2,S_k} \\ \vdots \\ c_{k,l_p,S_k} \end{bmatrix} = \begin{bmatrix} \langle \mathbf{f}_{l_1}, \mathbf{f}_k \rangle \\ \langle \mathbf{f}_{l_2}, \mathbf{f}_k \rangle \\ \vdots \\ \langle \mathbf{f}_{l_p}, \mathbf{f}_k \rangle \end{bmatrix}. \tag{43}$$

If the frame $\{\mathbf{f}_l | l \in S_k\}$ is redundant, the coefficients are not unique. One option for the solution above would be to use the pseudoinverse of the matrix. This is equivalent to computing the inner product of $\mathbf{f}_k$ with the dual frame of $\{\mathbf{f}_l | l \in S_k\}$ in $\mathcal{S}_k$, which we denote by $\{\phi_l^{S_k} | l \in S_k\}$: $c_{k,l,S_k} = \langle \mathbf{f}_k, \phi_l^{S_k} \rangle$. The

projection is equal to:

$$\mathcal{P}_{\mathcal{S}_k}(-e_k\mathbf{f}_k) = -e_k \sum_{l\in S_k} c_{k,l,S_k}\mathbf{f}_l. \tag{44}$$

Consistent with section 3, we change step 3 of the algorithm to:

3. Update $\{a_l | l \in S_k\}$ to $a'_l = a_l - e_k c_{k,l,S_k}$, where c_{k,l,S_k} satisfy (43).

Similarly, the residual is $-e_k\tilde{c}_{k,S_k}\mathbf{r}_{k,S_k}$, where

$$\tilde{c}_{k,S_k} = ||\mathbf{f}_k - \sum_{l\in S_k} c_{k,l,S_k}\mathbf{f}_l||, \text{ and} \tag{45}$$

$$\mathbf{r}_{k,S_k} = \frac{\mathbf{f}_k - \sum_{l\in S_k} c_{k,l,S_k}\mathbf{f}_l}{||\mathbf{f}_k - \sum_{l\in S_k} c_{k,l,S_k}\mathbf{f}_l||}. \tag{46}$$

This corresponds to expressing $e_k\mathbf{f}_k$ as the direct sum of the vectors $e_k\tilde{c}_{k,S_k}\mathbf{r}_{k,S_k} \oplus e_k \sum_{l\in S_k} c_{k,l,S}\mathbf{f}_l$, and compensating only for the second part of this sum. Note that $\tilde{c}_{k,S_k}$ and $\mathbf{r}_{k,S_k}$ are the same independent on whether we use the pseudoinverse to solve (43) or any other left inverse.

The modification to the equations for the total error and the corresponding cost functions are straightforward:

$$\mathcal{E} = \sum_{k=1}^{M} e_k\tilde{c}_{k,S_k}\mathbf{r}_{k,S_k} \tag{47}$$

$$E\{||\mathcal{E}||^2\} = \frac{\Delta^2}{12}\sum_{k=1}^{M}\tilde{c}^2_{k,S_k}, \text{ and} \tag{48}$$

$$||\mathcal{E}|| \leq \frac{\Delta}{2}\sum_{k=1}^{M}\tilde{c}_{k,S_k}. \tag{49}$$

When $S_k = \{l_k\}$ for $k < M$, this collapses to a tree quantizer. Similarly, when $S_k = \{k+1\}$, the structure becomes a sequential quantizer. Since the tree quantizers is a special case of the higher order quantizer, it is straightforward to show that for a given frame vector ordering a higher order quantizer can always achieve the cost of a tree quantizer. Note that S_M is always empty, and, therefore $\tilde{c}_{M,S_M} = ||\mathbf{f}_M||$, which is consistent with the cost analysis for the first order quantizers.

For appropriately ordered finite frames in N dimensions, the first $M-N$ error coefficients $\tilde{c}_{k,S_k}$ can be forced to zero with an N^{th} or higher order quantizer. In this case, the error coefficients determining the cost of the quantizer are the remaining N ones—the error becomes $\sum_{k=M-N+1}^{M} e_k\tilde{c}_{k,S_k}\mathbf{r}_{k,S_k}$, with the corresponding cost functions modified accordingly. One way to achieve that function is to use all the unquantized coefficients to compensate for the quantization of coefficient a_k by setting $S_k = \{(k+1),\ldots,M\}$ and ordering the vectors such that the last N frame vectors span the space. Another way to achieve this cost function is discussed as an example in next section.

Unfortunately, the design space for higher order quantizers is quite large. The optimal frame vector ordering and S_k selection is still an open question and we do not attempt it in this work.

7 Experimental Results

To validate the theoretical results we presented above, in this section we consider the same example as was included in [1, 2]. We use the tight frame consisting of the 7^{th} roots of unity to expand randomly selected vectors in $\mathbb{R}^2$, uniformly distributed inside the unit circle. The frame expansion is quantized using $\Delta = 1/4$, and the vectors are reconstructed using the corresponding synthesis frame. The frame vectors and the coefficients relevant to quantization are given by:

$$\underline{\mathbf{f}}_n = (\cos(2\pi n/7), \sin(2\pi n/7)), \tag{50}$$

$$\mathbf{f}_n = ((2/7)\cos(2\pi n/7), (2/7)\sin(2\pi n/7)), \tag{51}$$

$$c_{k,l} = \cos\left(2\pi(k-l)/7\right), \tag{52}$$

$$\tilde{c}_{k,l} = (2/7)|\sin\left(2\pi(k-l)/7\right)|. \tag{53}$$

For this frame the natural ordering is suboptimal given the criteria we propose. An optimal ordering of the frame vectors is $(\mathbf{f}_1, \mathbf{f}_4, \mathbf{f}_7, \mathbf{f}_3, \mathbf{f}_6, \mathbf{f}_2, \mathbf{f}_5)$, and we refer to it as such for the remainder of this section, in contrast to the natural frame vector ordering. A sequential quantizer with this optimal ordering meets the lower bound for the cost under both cost functions we propose. Thus, it is an optimal first order noise shaping quantizer for both cost functions. We compare this strategy to the one proposed in [1, 2] and also explored as a special case of section 6.1. Under that strategy, there is no projection performed, just error propagation. Therefore, based on the frame variation as described in [1, 2], the natural frame ordering is the best ordering to implement that strategy.

In the simulations, we also examine the performance of higher order quantization, as described in section 6.2. Since we operate on a two dimensional frame, a second order quantizer can perfectly compensate for the quantization of all but the last two expansion coefficients. Therefore, all the error coefficients of equation (47) are 0, except for the last two. A third order or higher quantizer should not be able to improve the quantization cost. However, the ordering of frame vectors is still important, since the angle between the last two frame vectors to be quantized affects the error, and should be as small as possible.

To visualize the results we plot the distribution of the reconstruction error magnitude. In figure 3(a) we consider the case of direct coefficient quantization. Figures 3(b) and (c) correspond to noise shaping using the natural and the optimal frame ordering respectively, and the method proposed in [1, 2], i.e. without projecting the error. Figures 3(d), (e), and (f) use the projection method we propose using the natural frame ordering, and first, second and third order projections, respectively. Finally, figures 3(g) and (h) demonstrate first and second order noise shaping results, respectively, using projections on the optimal frame ordering. For clarity of the legend we do not plot the third order results; they

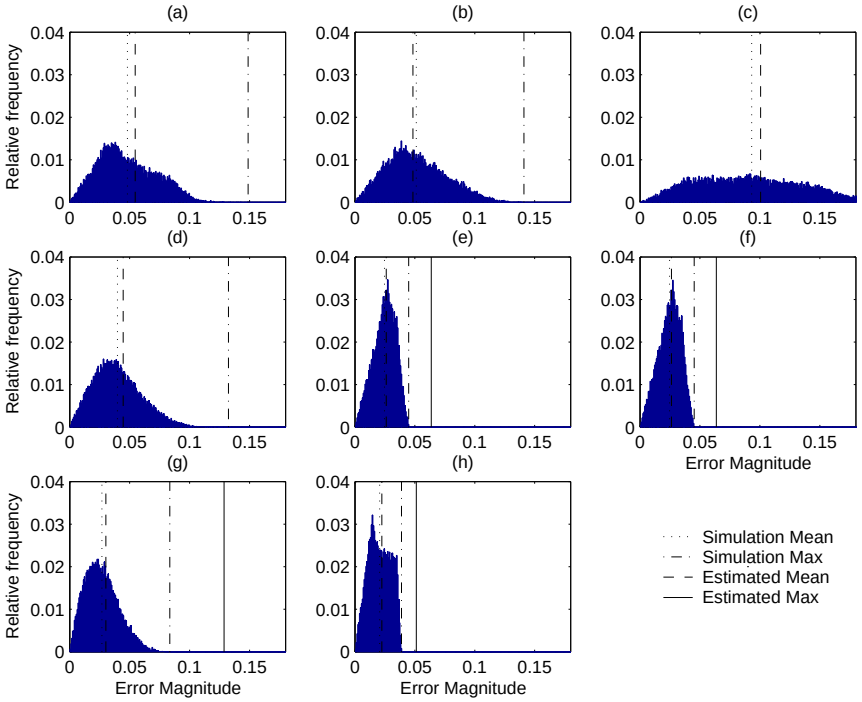

Figure 3: Histogram of the reconstruction error under (a) direct coefficient quantization, (b) natural ordering and error propagation without projections, (c) optimal ordering and error propagation without projections. In the second row, natural ordering using projections, with (d) first, (e) second, and (f) third order error propagation. In the third row, optimal ordering using projections, with (g) first and (h) second order error propagation (the third order results are similar to the second order ones but are not displayed for clarity of the legend).

are almost identical to the second order case. On all the plots we indicate with dotted and dash-dotted lines the average and maximum reconstruction error respectively, and with dashed and solid line the average and maximum error, as determined using the cost functions of section 4[3].

The results show that the projection method results in smaller error, even using the natural frame ordering. As expected, the results using the optimal frame vector ordering are the best among the simulations we performed. The simulations also confirm that in $\mathbb{R}^2$, noise shaping provides no benefit beyond second order and that the frame vector ordering affects the error even in higher order noise shaping, as predicted by the analysis. It is evident that the upper bound model is loose, as expected. The error average, on the other hand, is

[3]In some parts of the figure, the lines are out of the axis bounds. For completeness, we list the results here: (a) Estimated Max=0.25, (b) Estimated Max=0.22, (c) Estimated Max=0.45, Simulation Max=0.27, (d) Estimated Max=0.20.

surprisingly close to the simulation mean, although it usually overestimates it.

Our results were similar for a variety of frame expansions on different dimensions, redundancy values, vector orderings, and noise shaping orders, including oblique bases (i.e. non-redundant frame expansions), validating the theory developed in the previous sections.

8 Extensions to Infinite Frames

When extending the results above to frames with a countably infinite numbers of synthesis frame vectors, we let $M \to \infty$ and modify equations (22), (28), and (31) to reflect an error rate corresponding to average error per frame vector, or equivalently per expansion coefficient. As $M \to \infty$ the effect of the last term on the error rate tends to zero. Consequently in considering the error rate we replace equations (22), (28), and (31) by

$$\overline{\mathcal{E}} = \lim_{M\to\infty} \frac{1}{M} \sum_{k=0}^{M-1} e_k \tilde{c}_{k,k+1} \mathbf{r}_{k,k+1}, \tag{54}$$

$$\overline{E\left\{||\mathcal{E}||^2\right\}} = \lim_{M\to\infty} \frac{1}{M} \frac{\Delta^2}{12} \left(\sum_{k=0}^{M-1} \tilde{c}_{k,k+1}^2 \right), \text{ and} \tag{55}$$

$$\overline{||\mathcal{E}||} \leq \lim_{M\to\infty} \frac{1}{M} \frac{\Delta}{2} \left(\sum_{k=0}^{M-1} \tilde{c}_{k,k+1} \right), \tag{56}$$

respectively, where $\overline{(\cdot)}$ denotes rate, and the frame vectors are indexed in $\mathbb{N}$. Similar modifications are straightforward for the cases of tree[4] and higher order quantizers, and for any countably infinite indexing of the frame vectors. At the design stage, the choice of frame should be such to ensure convergence of the cost functions. In the remaining of this section we expand further on shift invariant frames, where convergence of the cost functions is straightforward to demonstrate.

8.1 Infinite Shift Invariant Frames

We define infinite shift invariant reconstruction frames as infinite frames $\mathbf{f}_k$ for which the inner product between frame vectors $\langle \mathbf{f}_k, \mathbf{f}_l \rangle$ is a function only of the index difference $k-l$. Consistent with traditional signal processing terminology we define this as the autocorrelation of the frame: $R_m = \langle \mathbf{f}_k, \mathbf{f}_{k+m} \rangle$. Shift invariance implies that the reconstruction frame is uniform, with $||\mathbf{f}_k||^2 = \langle \mathbf{f}_k, \mathbf{f}_k \rangle = R_0$.

An example of such a frame is an LTI system: consider a signal $x[n]$ that is quantized to $\hat{x}[n]$ and filtered to produce $\hat{y}[n] = \sum_k \hat{x}[k] h[n-k]$. We consider the coefficients $x[k]$ to be a frame expansion of $y[n]$, where $h[n-k]$ are the

[4] This is a slight abuse of the term, since the resulting infinite graph might have no root.

reconstruction frame vectors $\mathbf{f}_k$. We rewrite the convolution equation as:

$$y[n] = \sum_k x[k]h[n-k] = \sum_k x[k]\mathbf{f}_k[n], \tag{57}$$

where $\mathbf{f}_k[n] = h[n-k]$. Equivalently, we may consider $x[n]$ to be quantized, converted to continuous time impulses, and then filtered to produce $\hat{y}(t) = \sum_k \hat{x}[k]h(t-kT)$. We desire to minimize the quantization cost after filtering, compared to the signals $y[n] = \sum_k x[k]h[n-k]$ and $y(t) = \sum_k x[k]h(t-kT)$, assuming the cost functions we described.

For the remainder of this section we only discuss the discrete-time version of the problem since the continuous time development is identical. The corresponding frame autocorrelation functions are $R_m = R_{hh}[m] = \sum_m h[n]h[n-m]$ in the discrete-time case and $R_m = R_{hh}(mT) = \int h(t)h(t-mT)dt$ in the continuous-time case. A special case of this setup is the oversampling frame, in which $h(t)$ or $h[n]$ is the ideal lowpass filter used for the reconstruction, and $R_m = \mathrm{sinc}(\pi m/r)$, where r is the oversampling ratio.

8.2 First Order Noise Shaping

Given a shift invariant frame, it is straightforward to determine the coefficients $c_{k,l}$ and $\tilde{c}_{k,l}$ that are important for the design of a first order quantizer. These coefficients are also shift invariant, so we denote them using $c_m = c_{k,k+m}$ and $\tilde{c}_m = \tilde{c}_{k,k+m}$. Combining equations (19) and (21) from section 3 and the definition of R_m above, we compute the relevant coefficients:

$$c_m = c_{-m} = \frac{R_m}{R_0} \tag{58}$$

$$\tilde{c}_m = \tilde{c}_{-m} = \sqrt{R_0(1-c_m^2)} \tag{59}$$

For every coefficient a_k of the frame expansion and corresponding frame vector $\mathbf{f}_k$, the vector that minimizes the projection error is the vector $\mathbf{f}_{k\pm m_o}$, where $m_o > 0$ minimizes $\tilde{c}_m$, or, equivalently, maximizes $|c_m|$, i.e. $|R_m|$. By symmetry, for any such m_o, $-m_o$ is also a minimum. Due to the shift invariance of the frame, m_o is the same for all frame vectors. Projecting to $\mathbf{f}_{k+m_o}$ or $\mathbf{f}_{k-m_o}$ generates a path with no loops, and therefore the optimal tree quantizer path, as long as the direction is consistent for all the coefficients. When $m_o = 1$, the optimal tree quantizer is also an optimal sequential quantizer. The optimality holds under both the additive noise model and the error upper bound model.

In the case of filtering, the noise shaping implementation is shown in figure 4, with $H_f(z) = c_{m_o} z^{-m_o}$. It is easy to show that for the special case of the oversampling frame $m_o = 1$, confirming that the time sequential ordering of the frame vectors is optimal for the given frame.

8.3 Higher Order Noise Shaping

As we discuss in section 6.2, determining the optimal ordering for higher order quantization is not straightforward. Therefore, in this section we consider higher

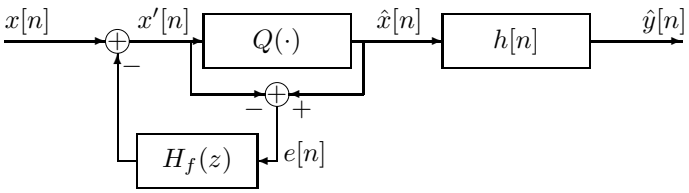

Figure 4: Noise shaping quantizer, followed by filtering

order noise shaping for the natural frame ordering, assuming that when a_k is quantized, the next p coefficients, $a_{k+1}, \ldots, a_{k+p}$, are used for compensation by updating them to

$$a'_{k+l} = a_{k+l} - e_k c_l, \ l = 1, \ldots, p. \tag{60}$$

The coefficients c_l project $\mathbf{f}_k$ onto the space $\mathcal{S}_k$ defined by $\{\mathbf{f}_{k+1}, \ldots, \mathbf{f}_{k+p}\}$. Because of the shift invariance property, these coefficients are independent of k. Shift invariance also simplifies equation (43):

$$\begin{bmatrix} R_0 & R_1 & \cdots & R_{p-1} \\ R_1 & R_0 & \cdots & R_{p-2} \\ \vdots & & \ddots & \vdots \\ R_{p-1} & & \cdots & R_0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_p \end{bmatrix}, \tag{61}$$

with R_m being the frame autocorrelation function. There are several options for solving this equation, including the Levinson recursion.

The implementation for higher order noise shaping before filtering is shown in figure 4, with $H_f(z) = \sum_{l=1}^{p} c_l z^{-l}$, where the c_l solve (61). The feedback filter implements the projection and the coefficient update described in equation (60).

For the special case of the oversampling frame, table 1 demonstrates the benefit of adjusting the feedback loop to perform a projection. The table reports the approximate dB gain in reconstruction error energy using the solution to (61) compared to the classical feedback loop implied by (41). For example, for oversampling ratios greater than 8 and third order noise shaping, there is an 8dB gain in implementing the projection method. The gain figures in the table are calculated using the additive noise model of quantization.

	$r=2$	$r=4$	$r=8$	$r=16$	$r=32$	$r=64$
$p=1$	0.9	0.2	0.1	0.0	0.0	0.0
$p=2$	4.5	3.8	3.6	3.5	3.5	3.5
$p=3$	9.1	8.2	8.0	8.0	8.0	8.0
$p=4$	14.0	13.1	12.9	12.8	12.8	12.8

Table 1: Gain in dB in in-band noise power comparing p^{th} order classical noise shaping with p^{th} order noise shaping using projections.

The applications in this section can be extended for frames generated by oversampled filterbanks, a case extensively studied in [3]. In that work, the

problem is posed in terms of prediction with quantization of the prediction error. Motivated by that work, we determined the solution to the filterbank problem using the projective approach. Setting up and solving for the compensation coefficients using equation (43) in section 6.2 corresponds exactly to solving equation (21) in [3], the solution to that setup under the white noise assumption.

It is reassuring that our approach, although different from [3] generates the same solution. Conveniently, the experimental results from that work apply in our case as well. Our theoretical results complement [3] by providing a projective viewpoint to the problem, developing a deterministic cost function and showing that even in the case of critically sampled biorthogonal filterbanks noise shaping can provide improvements compared to scalar coefficient quantization. On the other hand, it is not straightforward to use our approach to analyze and compensate for colored additive noise, as described in [3].

Acknowledgment

We express our thanks to the anonymous reviewers for their insightful and helpful comments during the review process.

References

[1] J. J. Benedetto, A. M. Powell, and O. Yilmaz. Sigma-delta quantization and finite frames. *IEEE Transactions on Information Theory*, submitted for publication. Available: `http://www.math.umd.edu/~jjb/ffsd.pdf`.

[2] J.J. Benedetto, O. Yilmaz, and A.M. Powell. Sigma-delta quantization and finite frames. In *Proceedings of IEEE ICASSP 2004*, Montreal, Canada, May 2004. IEEE.

[3] H. Bolcskei and F. Hlawatsch. Noise reduction in oversampled filter banks using predictive quantization. *IEEE Transactions on Information Theory*, 47(1):155–172, Jan 2001.

[4] P. Boufounos and A.V. Oppenheim. Quantization noise shaping on arbitrary frame expansions. In *Proceedings of IEEE ICASSP 2005*, Philadelphia, PA, USA, March 2005. IEEE.

[5] J.C. Candy and G.C Temes, editors. *Oversampling Delta-Sigma Converters*. IEEE Press, 1992.

[6] T.H. Cormen, C.E. Leiserson, and R.L. Rivest. *Introduction to algorithms*. MIT Press and McGraw-Hill, 2nd edition, 2001.

[7] Z. Cvetkovic and M. Vetterli. Overcomplete expansions and robustness. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, 1996.*, pages 325–328, Jun 1996.

[8] I. Daubecies. *Ten Lectures on Wavelets*. CBMS-NSF regional conference series in applied mathematics. SIAM, Philadelphia, PA, 1992.

[9] V.K. Goyal, M. Vetterli, and N.T. Thao. Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms. *IEEE Transactions on Information Theory*, 44(1):16–31, Jan 1998.

[10] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, 1996.

[11] N.T. Thao and M. Vetterli. Reduction of the MSE in R-times oversampled A/D conversion $O(1/R)$ to $O(1/R^2)$. *IEEE Transactions on Signal Processing*, 42(1):200–2003, Jan 1994.