CONDITIONS FOR OPTIMALITY OF SCALAR FEEDBACK QUANTIZATION

Milan S. Derpich, Daniel E. Quevedo, and Graham C. Goodwin

School of Electrical Engineering and Computer Science, The University of Newcastle, NSW, Australia

ABSTRACT

This paper presents novel results on *scalar feedback quantization* (SFQ) with uniform quantizers. We focus on general SFQ configurations where reconstruction is via a linear combination of frame vectors. Using a deterministic approach, we derive two necessary and sufficient conditions for SFQ to be optimal, i.e., to produce, for every input, a quantized sequence that is a global minimizer of the 2-norm of the reconstruction error. The first optimality condition is related to the design of the feedback quantizer, and can always be achieved. The second condition depends only on the reconstruction vectors, and is given explicitly in terms of the Gram matrix of the reconstruction frame. As a by-product, we also show that the the first condition alone characterizes scalar feedback quantizers that yield the smallest MSE, when one models quantization noise as uncorrelated, identically distributed random variables.

***Index Terms*—** Frames, Quantization, Sigma-Delta Modulation.

1. INTRODUCTION

In many signal processing applications, signals have to be represented by a series of numbers (samples), so that they can be processed, transmitted or stored in digital form. This paradigm requires sampling, quantization and reconstruction.

The quantization of the samples, namely the sequence $\{c_j\}_{j=1}^N$, $N \in \mathbb{N}$, yields a sequence $\{\mu_j\}_{j=1}^N$ whose elements are constrained to belong to a discrete set of scalars. We focus our attention on uniform quantization, and thus require that

$$\mu_j \in \mathbb{U}, \forall j \in \{1, ..., N\}; \quad \mathbb{U} \triangleq \{k\Delta : k \in \mathbb{Z}, \Delta \in \mathbb{R}^+\}. \quad (1)$$

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

The simplest and most common paradigm to recover the signal from the numbers is *linear reconstruction*. Here, one is able to recover the original signal, say $\mathbf{a}$, via

$$\mathbf{a} = \sum\nolimits_{j=1}^{N} c_j \psi_j, \quad (2)$$

In (2), $\{\psi_j\}_{j=1}^N$ is a set of vectors (a frame) in the reconstruction Hilbert space $\mathcal{W}$ (typically a subspace of ℓ^2 or of L^2). Thus, the samples $\{c_j\}_{j=1}^N$ are the frame expansion coefficients of $\mathbf{a}$. Examples of linear reconstruction are the Shannon-Whittaker reconstruction formula, the reconstruction stage in filter-banks, and the inverse wavelet-transform.

Throughout this work, we will be concerned with the squared 2-norm of the reconstruction error, i.e.,

$$D(\mathbf{c}, \boldsymbol{\mu}) \triangleq \left\| \mathbf{a} - \sum\nolimits_{j=1}^{N} \mu_j \psi_j \right\|^2 = \left\| \sum\nolimits_{j=1}^{N} (\mu_j - c_j)\psi_j \right\|^2, \quad (3)$$

where $\|\cdot\|^2 = \langle \cdot, \cdot \rangle_{\mathcal{W}}$ (the latter being the inner product in $\mathcal{W}$).

An optimal *vector quantizer* [1], which jointly quantizes the entire sequence $\mathbf{c}$, yields the *minimum achievable distortion*

$$D^\star(\mathbf{c}) \triangleq \min_{\boldsymbol{\mu} \in \mathbb{U}^N} D(\mathbf{c}, \boldsymbol{\mu}), \quad (4)$$

for every $\mathbf{c} \in \mathbb{R}^N$. Unfortunately, minimization of (3) subject to (1) is a non-convex optimization problem. Moreover, the complexity of solving this problem grows exponentially with the number of coefficients to be quantized. In addition, unless $\{\psi_j\}_{j-1}^N$ forms an orthogonal set, one would need to "preview" the entire input sequence before being able to calculate any optimal quantized value for μ_j. This is incompatible with delay sensitive applications.

For the above reasons, in practice quantization is often accomplished via simpler sub-optimal methods that operate sequentially. The simplest of these correspond to *scalar feedback* (SF) quantizers. At the i-th iteration, these A/D converters obtain the output sample[1] u_i by simple scalar quantization of an auxiliary sequence, which is a linear combination of input and output samples, i.e.,

$$u_i = \mathcal{Q}\Big(\sum\nolimits_{j:j\le i} \alpha_{i,j} c_j + \sum\nolimits_{j:j<i} \beta_{i,j}(u_j - c_j) \Big). \quad (5a)$$

In (5), the real scalars $\alpha_{i,j}, \beta_{i,j}$, $i, j \in \{1, 2, \ldots, N\}$ are design parameters, and $\mathcal{Q}(\cdot)$ is the *nearest neighbour scalar quantization function*

$$\mathcal{Q}(v) \triangleq \arg\min_{\mu \in \mathbb{U}} |v - \mu|, \quad \forall v \in \mathbb{R}. \quad (5b)$$

The above expressions can be used to describe many scalar quantization schemes, including PCM, DPCM and (multi-bit) Sigma-Delta ($\Sigma\Delta$) converters [2]. The latter have been well studied in the context of shift-invariant reconstruction spaces (wherein reconstruction is done by LTI filters), and recently also for frame expansions (see, e.g., [3, 4]).

Not surprisingly, for a given reconstruction frame, and in return for the above mentioned shortcomings, optimal vector quantization generally outperforms SFQ. However, it is not known under what conditions this performance gap exists. In this paper we derive those conditions. More precisely, we state necessary and sufficient conditions for SFQ to be optimal, i.e., to yield, for any input $\mathbf{c}$, the quantized output sequence $\boldsymbol{\mu}$ that minimizes $D(\mathbf{c}, \boldsymbol{\mu})$ in (3). Our results extend the work documented in [5, 6] to more general situations.

Notation We use bold lowercase letters, e.g. $\mathbf{x}$, to denote both the sequence $\{x_j\}_{j=1}^N$ and the column vector $[x_1 \cdots x_N]^T$, where the meaning is clear from the context. We also use bold letters to represent matrices (uppercase) and their corresponding column vectors (lowercase). For example, if $\mathbf{G}$ is a matrix, we use $\mathbf{g}_i$ to refer to the i-th column of $\mathbf{G}$, and $g_{i,j}$ to refer to the j-th element of $\mathbf{g}_i$. The *null space* and the *Moore-Penrose pseudo-inverse* of a matrix $\mathbf{G}$ are

E-mails: milan.derpich@studentmail.newcastle.edu; dquevedo@ieee.org; graham.goodwin@newcastle.edu.au

[1] Hereafter we use μ to denote arbitrary quantized values and reserve the symbol u for the output of the SF quantizer.

denoted respectively via $\mathcal{N}(\mathbf{G})$ and $\mathbf{G}^\dagger$. The notation $\mathbf{G}^i$ refers to the sub-matrix obtained by removing the first i columns and i rows from $\mathbf{G}$. Similarly, $\mathbf{x}^j$ denotes the vector $\mathbf{x}$ without its first j elements. The symbol $\mathbf{0}_N$ denotes an N-length column vector of zeros. We use the short-hand notation $\|\mathbf{x}\|^2_{\mathbf{G}}$ for the quadratic form $\mathbf{x}^T\mathbf{G}\mathbf{x}$. We write "iff" as an abbreviation for "if and only if". $\mathbb{Z}$ corresponds to the integers, and we use $\mathbb{Z}^N$ to denote the set of all N-length vectors with integer elements. We say a matrix or vector is *integral* iff all its elements are integers.

2. PRELIMINARIES

2.1. Brief Overview of Frames

Here we will first present some facts regarding frames that will be used in our subsequent analysis[2].

A finite frame for a Hilbert space $\mathcal{W}$ is an ordered set of vectors $\{\psi_j\}_{j=1}^N \subset \mathcal{W}$ such that, for every $\mathbf{w} \in \mathcal{W}$,

$$A\|\mathbf{w}\|^2 \leq \sum\nolimits_{j=1}^{N} |\langle \mathbf{w}, \psi_j \rangle_{\mathcal{W}}|^2 \leq B\|\mathbf{w}\|^2, \tag{6}$$

where the scalar constants A, B satisfy $0 < A \leq B < \infty$.

The *synthesis operator* $\Psi : \ell^2 \to \mathcal{W}$ of the frame $\{\psi_j\}_{j=1}^N$ is defined via

$$\Psi\mathbf{c} = \sum\nolimits_{j=1}^{N} c_j \psi_j, \quad \forall \mathbf{c} \in \ell^2, \tag{7}$$

where ℓ^2 denotes the set of square-summable sequences. The *Gram matrix* $\mathbf{G} \in \mathbb{R}^{N\times N}$ of the frame $\{\psi_j\}_{j=1}^N$ is defined element-wise via

$$G_{j,k} \triangleq \langle \psi_j, \psi_k \rangle, \;\; j,k \in \{1,\ldots,N\}. \tag{8}$$

It thus follows that

$$\langle \Psi\mathbf{x}, \Psi\mathbf{y} \rangle_{\mathcal{W}} = \mathbf{x}^T\mathbf{G}\mathbf{y}, \quad \forall \mathbf{x}, \mathbf{y} \in \ell^2 \tag{9}$$

which implies that $\mathbf{G}$ is positive semi-definite, and, in particular, that

$$\|\Psi\mathbf{e}\|^2_{\mathcal{W}} = \|\mathbf{e}\|^2_{\mathbf{G}}, \quad \forall \mathbf{e} \in \ell^2. \tag{10}$$

2.2. Feedback Quantization of Frame Expansions

It is easy to show from (5) that an SF quantizer cannot yield $D(\mathbf{c},\mathbf{u}) = D^\star(\mathbf{c})$ for all $\mathbf{c} \in \mathbb{R}^N$ unless[3] $\alpha_{i,j} = \delta_{i,j}$, $\forall i,j$, where $\delta_{i,j}$ is the Kronecker delta function. If the latter holds, then (5) can be written as follows:

$$\mathbf{u} \triangleq \mathcal{Q}(\mathbf{v}); \qquad \mathbf{v} \triangleq \mathbf{c} - \mathbf{F}\mathbf{n}; \qquad \mathbf{n} \triangleq \mathbf{u} - \mathbf{v}, \tag{11}$$

where $\mathcal{Q}(\mathbf{v}) = [\mathcal{Q}(v_1) \cdots, \mathcal{Q}(v_N)]^T$, $\mathbf{F}$ is the *feedback matrix* and $\mathbf{n}$ is the vector of *quantization errors*. In order for the above equations to be well defined, $\mathbf{F}$ needs to be lower *strictly-triangular*, i.e., lower triangular with all main diagonal elements equal to zero[4]. Notice also that $\mathbf{F}$ is the only degree of freedom in the design of an SF quantizer.

In order to determine $D(\mathbf{c},\mathbf{u})$ for SF quantizers, it is convenient to define the *noise shaping matrix*

$$\mathbf{S} \triangleq (\mathbf{I}_N - \mathbf{F}), \tag{12}$$

where $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. Clearly, $\mathbf{S}$ is constrained to be lower *unit-triangular*, i.e., lower triangular with all its main diagonal elements equal to 1.

[2] A deeper treatment of the subject can be found, e.g., in [7].

[3] To verify this, it suffices to consider an input sequence $\mathbf{c} \in \mathbb{U}^N$ in (5a).

[4] Otherwise (11) cannot be solved sequentially.

Substituting (12) into (11) yields $\mathbf{u} = \mathbf{c} + \mathbf{S}\mathbf{n}$. Using this, and substituting (10) and (7) into (3), the distortion achieved by SFQ can be written as

$$D(\mathbf{c},\mathbf{u}) = \|\Psi(\mathbf{u}-\mathbf{c})\|^2_{\mathcal{W}} = \|\mathbf{u}-\mathbf{c}\|^2_{\mathbf{G}} = \mathbf{n}^T\mathbf{S}^T\mathbf{G}\,\mathbf{S}\,\mathbf{n}. \tag{13}$$

3. MAIN RESULT

We can now state the main result of this paper.

Theorem 1 *For any given reconstruction frame with Gram matrix $\mathbf{G} \in \mathbb{R}^{N\times N}$, the distortion $D(\mathbf{c},\mathbf{u})$ of an SF quantizer equals $D^\star(\mathbf{c})$ for all $\mathbf{c} \in \mathbb{R}^N$ iff the following two conditions hold:*

(i) The columns of the associated feedback matrix $\mathbf{F}$ satisfy

$$\mathbf{f}^i_i = \mathbf{m}_i + \boldsymbol{\varsigma}_i, \qquad \forall i \in \{1,\ldots,N\}, \tag{14a}$$

$$\textit{where } \mathbf{m}_i \triangleq {\mathbf{G}^i}^\dagger \mathbf{g}^i_i, \qquad \forall i \in \{1,\ldots,N\}, \tag{14b}$$

and where the vectors $\{\boldsymbol{\varsigma}_i\}_{i=1}^N$ satisfy $\boldsymbol{\varsigma}_i \in \mathcal{N}(\mathbf{G}^i)$, but are otherwise arbitrary.

(ii) For every $i \in \{1,\ldots,N\}$, $\exists \boldsymbol{\xi}_i \in \mathcal{N}(\mathbf{G}^i)$ such that

$$\mathbf{m}_i + \boldsymbol{\xi}_i \in \mathbb{Z}^{N-i}, \tag{15}$$

i.e., such that $\mathbf{m}_i + \boldsymbol{\xi}_i$ is an integral vector. ▲

Notice that (i) describes a "matching" condition between the feedback matrix and the reconstruction frame. Thus, (i) can always be satisfied by a proper choice of $\mathbf{F}$ (which is given explicitly by (14)). On the other hand, condition (ii) depends only on the reconstruction frame, or more precisely, on its Gram matrix.

The proof of Theorem 1 will be given in Section 6, based on preliminary results given in Sections 4 and 5. The latter provide valuable insight into the SFQ problem, and stem from two alternative approaches: lattice quantization and dynamic programming.

4. LATTICE QUANTIZATION FORMULATION

In this section we use the fact that minimization of (3) subject to (1) is equivalent to a lattice quantization problem. To show this, we first note that any symmetric positive semidefinite matrix $\mathbf{G} \in \mathbb{R}^{N\times N}$ can be decomposed as

$$\mathbf{G} = \mathbf{H}^T\mathbf{H}, \tag{16}$$

where $\mathbf{H} \in \mathbb{R}^{N\times N}$ is lower triangular (see., e.g., [8]). It then follows directly from (9) and (16) that $\langle \Psi\mathbf{x}, \Psi\mathbf{y} \rangle_{\mathcal{W}} = \langle \mathbf{H}\mathbf{x}, \mathbf{H}\mathbf{y} \rangle$, $\forall \mathbf{x}, \mathbf{y} \in \ell^2$. Thus, one can analyze the relationship between the images in $\mathcal{W}$ through Ψ of any group of sequences by looking at their images in ℓ^2 through $\mathbf{H}$. In particular,

$$D(\mathbf{c},\boldsymbol{\mu}) = \|\Psi(\boldsymbol{\mu}-\mathbf{c})\|^2_{\mathcal{W}} = \|\mathbf{H}\boldsymbol{\mu} - \mathbf{H}\mathbf{c}\|^2 = \|\boldsymbol{\mu}-\mathbf{c}\|^2_{\mathbf{G}}, \tag{17}$$

see (13) and (8).

Since the quantization alphabet $\mathbb{U}$ is uniform (see (1)), the images through $\mathbf{H}$ of all the sequences $\boldsymbol{\mu} \in \mathbb{U}^N$ constitute the *reconstruction lattice*

$$\mathcal{L} \triangleq \{\mathbf{H}\boldsymbol{\mu} \,:\, \boldsymbol{\mu} \in \mathbb{U}^N\} = \mathbf{H}\mathbb{Z}^N\Delta. \tag{18}$$

Accordingly, we say that $\mathbf{H}\Delta$ is the *generating matrix* for $\mathcal{L}$. Every lattice has a basic *Voronoi cell*, $\mathcal{V}_0$, associated with it, i.e., the region of points closer to the origin than to any other point in the lattice. More precisely,

$$\mathcal{V}_0 \triangleq \{\mathbf{x} = \mathbf{H}\mathbf{c} \,:\, \mathbf{c} \in \mathbb{R}^N,\; \|\mathbf{x}\| \leq \|\mathbf{x}-\boldsymbol{\beta}\|,\; \forall \boldsymbol{\beta} \in \mathcal{L}\}. \tag{19}$$

The Voronoi region around a lattice point $\mathbf{H}\boldsymbol{\mu} \in \mathcal{L}$ is the region $\mathcal{V}(\mathbf{H}\boldsymbol{\mu}) \triangleq \mathbf{H}\boldsymbol{\mu} + \mathcal{V}_0$. Similarly, we define the *quantization cell* around $\mathbf{H}\boldsymbol{\mu} \in \mathcal{L}$ of an SFQ converter as

$$\mathcal{C}(\mathbf{H}\boldsymbol{\mu}) \triangleq \mathbf{H}\boldsymbol{\mu} + \mathbf{H}\mathbf{S}\mathcal{Y}, \tag{20}$$

where the hyper-cube $\mathcal{Y} \triangleq \{\boldsymbol{\eta} : \eta_j \in [-\frac{\Delta}{2}, \frac{\Delta}{2}], \forall j \in \{1, \ldots, N\}\}$ is the set containing all possible[5] quantization noise sequences. Thus, $\mathcal{C}(\mathbf{H}\boldsymbol{\mu})$ is the set of all target points $\mathbf{H}\mathbf{c}$ for which an SF quantizer outputs the sequence $\boldsymbol{\mu}$.

With the above definitions, we can now prove the necessity of condition (i) of Theorem 1.

Lemma 1 *For a reconstruction frame with gram matrix* $\mathbf{G}$*, the distortion* $D(\mathbf{c}, \mathbf{u})$ *of an SF quantizer can equal* $D^\star(\mathbf{c})$ *for all* $\mathbf{c} \in \mathbb{R}^N$ *only if condition (i) in Theorem 1 holds.* ▲

Proof 1 *In view of* (17) *and* (19)*, an SF quantizer is optimal (i.e.,* $D(\mathbf{c}, \mathbf{u}) = D^\star(\mathbf{c}), \forall \mathbf{c} \in \mathbb{R}^N$*) iff* $\mathcal{C}(\mathbf{0}_N) = \mathcal{V}_0$*. A key property of* $\mathcal{V}_0$ *is that it has the minimum second moment among all the cells whose* $\mathcal{L}$*-translations form a tessellation*[6]*, see [1]. Thus, an SF quantizer is a candidate to be optimal only if its matrix* $\mathbf{F}$ *minimizes the second moment of* $\mathcal{C}(\mathbf{0}_N)$*. This second moment can be readily shown to be given by* $\frac{\Delta^{N+2}}{12}$ *trace*$\{\mathbf{S}^T\mathbf{G}\mathbf{S}\} = \frac{\Delta^{N+2}}{12}$*trace*$\{(\mathbf{I} - \mathbf{F})^T\mathbf{G}(\mathbf{I} - \mathbf{F})\}$*. By using* (16)*, the* i*-th element of the trace can be written as* $\|\mathbf{H}[\mathbf{0}_{i-1}^T \quad 1 \quad -\mathbf{f}_i^{iT}]^T\|^2 = (H_{i,i})^2 + \|\mathbf{h}_i^i - \mathbf{H}^i\mathbf{f}_i^i\|^2$*, since* $\mathbf{H}$ *is lower triangular and* $\mathbf{F}$ *is lower strictly-triangular. The fact that each trace term depends only on its corresponding column of* $\mathbf{F}$ *implies that the trace is minimized iff each* $\mathbf{f}_i^i$ *minimizes* $\|\mathbf{h}_i^i - \mathbf{H}^i\mathbf{f}_i^i\|^2$*. Clearly, this happens iff*

$$\mathbf{f}_i^i = \mathbf{H}^{i\dagger}\mathbf{h}_i^i + \boldsymbol{\varsigma}_i, \ \forall i \in \{1, \ldots, N\}, \tag{21}$$

where $\boldsymbol{\varsigma}_i$ *is an arbitrary vector in* $\mathcal{N}(\mathbf{H}^i)$ *(and, thus, in* $\mathcal{N}(\mathbf{G}^i)$ *as well). Substitution of the identity* $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^\dagger\mathbf{A}^T$ *into* (21) *yields* (14)*, thus completing the proof.* □

Remark 1 *If one models* $\mathbf{n}$ *as a vector of uncorrelated, uniformly distributed (u.u.d), random variables, one gets* $MSE = \frac{\Delta^{N+2}}{12}$ *trace*$\{\mathbf{S}^T\mathbf{G}\mathbf{S}\} = \frac{\Delta^{N+2}}{12}$*trace*$\{(\mathbf{I} - \mathbf{F})^T\mathbf{G}(\mathbf{I} - \mathbf{F})\}$*. On the other hand, an SF quantizer whose feedback matrix satisfies* (21) *happens to characterize one of the noise shaping quantizers for frame expansions proposed in [3]. More precisely, condition (i) is satisfied by the variant in which the error associated with each quantized coefficient is projected onto* all *the coefficients ahead of the current iteration coefficient. Thus, Lemma 1 also shows that, using an u.u.d model for quantization errors, the latter scheme achieves the minimum MSE among all SF quantizers*[7]*.* ▲

5. DYNAMIC PROGRAMMING FORMULATION

Sequential quantization methods, such as SFQ, decide upon the value of each output coefficient sequentially. Insight can be gained by analyzing them from a dynamic programming point of view. The key point is that each of the decisions contributes additively to the cost defined in (3), leaving, after each step, a sub-problem similar in form to the original one. In turn, each of these sub-problems is determined by the decisions already made. The following result allows us to formalize these observations

Lemma 2 (Cost Decomposition) *Let* $\mathbf{G} \in \mathbb{R}^{N\times N}$ *be a positive semi-definite, symmetric matrix. Then,* $\forall i \in \{1, 2, \ldots, N\}$*, and* $\forall \mathbf{x}, \mathbf{t} \in \mathbb{R}^N$*, the following holds:*

$$\|\mathbf{x}^{i-1} - \mathbf{t}^{i-1}\|^2_{\mathbf{G}^{i-1}} = K_i(x_i - t_i)^2 + \|\mathbf{x}^i - \mathbf{t}^i + \mathbf{f}_i^i(x_i - t_i)\|^2_{\mathbf{G}^i},$$

where the scalars K_i *are defined as*

$$K_i \triangleq G_{i,i} - \mathbf{g}_i^{iT}\mathbf{G}^i\mathbf{g}_i^i, \quad \forall i \in \{1, \ldots, N\}, \tag{22}$$

and where the vectors $\{\mathbf{f}_i^i\}_{i=1}^N$ *satisfy* (14)*.* ▲

Proof 2 *The result follows from direct algebraic manipulation, using the identity* $\mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}^\dagger$ *and from the fact that* $\mathbf{g}_i^{iT} = \mathbf{g}_i^{iT}\mathbf{G}^{i\dagger}\mathbf{G}^i$*,* $\forall i \in \{1, \ldots, N\}$ *(which stems from* $\mathbf{G}$ *being positive semidefinite).* □

Recursive application of Lemma 2 to (17) allows one to split the total cost $D(\mathbf{c}, \boldsymbol{\mu})$ as follows:

$$D(\mathbf{c}, \boldsymbol{\mu}) = \|\boldsymbol{\mu} - \mathbf{t}_0\|^2_{\mathbf{G}} = \sum\nolimits_{j=1}^{i-1} K_j\eta_j^2 + \|\boldsymbol{\mu}^i - \mathbf{t}_i^i\|^2_{\mathbf{G}^i}, \tag{23}$$

see (17), where the scalars K_j are defined in (22), and where

$$\eta_j \triangleq \mu_j - t_{j,j}; \qquad \mathbf{t}_j \triangleq \mathbf{t}_{j-1} - \mathbf{f}_{j-1}\eta_{j-1}, \qquad \forall j, \tag{24}$$

with $\mathbf{t}_0 \triangleq \mathbf{c}$.

The summation on the right hand side of (23) represents the (irreducible) reconstruction error stemming from the first $i-1$ decisions. The *cost-to-go* after decision $i-1$ is the last term in (23). It has the same form as the original cost, but it contains the *updated* target vector $\mathbf{t}_i^i$. The latter can be regarded as a state vector which summarizes the effect of $\mathbf{c}^i$, and of previous decisions, on the cost-to-go.

6. PROOF OF THEOREM 1

Joint Sufficiency of (i) and (ii) It is well known in lattice theory that any two lattices $\mathcal{L}_1 = \mathbf{M}_1\mathbb{Z}^N$ and $\mathcal{L}_2 = \mathbf{M}_2\mathbb{Z}^N$, with $\mathbf{M}_1$ and $\mathbf{M}_2$ non-singular, are equal iff there exists an integral matrix $\mathbf{T}$ with $\det\{\mathbf{T}\} = \pm 1$ such that $\mathbf{M}_1 = \mathbf{T}\mathbf{M}_2$ (see, e.g., [9]). On the other hand, if conditions (i) and (ii) hold, then there exists a lower strictly-triangular matrix $\Xi \in \mathcal{N}(\mathbf{H})$ such that $\mathbf{S} + \Xi$ is integral. Since $\mathbf{S}$ is lower unit-triangular, we have that $\det\{S\} = \det\{\mathbf{S} + \Xi\} = 1$, and thus $\mathbb{Z}^N = (\mathbf{S} + \Xi)\mathbb{Z}^N$. It then follows that $\mathcal{L} = \mathbf{H}\mathbb{Z}^N = \mathbf{H}(\mathbf{S} + \Xi)\mathbb{Z}^N = \mathbf{H}\mathbf{S}\mathbb{Z}^N$. On the other hand, if condition (i) holds, then it follows directly from (21) that the product $\mathbf{H}\mathbf{S}$ yields an orthogonal, lower triangular matrix. This in turn implies that $\mathcal{L}$ is a rectangular lattice. Moreover, it is easy to verify that the associated Voronoi cell $\mathcal{V}_0$ is given by the hyper-rectangle $\mathbf{H}\mathbf{S}\mathcal{Y}$, which is precisely the quantization cell of the SF quantizer, $\mathcal{C}(\mathbf{0}_N)$ (see (20)). Therefore (i) and (ii) guarantee that the corresponding SF quantizer is optimal.

Necessity of Conditions (i) and (ii) The necessity of (i) was shown in Lemma 1. Thus, it suffices to prove the necessity of (ii) assuming that (i) holds. If (i) holds, then the target vectors $\mathbf{t}_j$ given in (24) can be written in terms of the feedback matrix $\mathbf{F}$ as follows

$$\mathbf{t}_j = \mathbf{c} - \mathbf{F}\begin{bmatrix} \mathbf{I}_{j-1} & \mathbf{0}_{N-j+1} \\ \mathbf{0}_{N-j+1} & \mathbf{0}_{N-j+1} \end{bmatrix}\boldsymbol{\eta}. \tag{25}$$

[5] Assuming the set of possible input sequences $\mathbf{c}$ is compact and big enough for its image through $\mathbf{H}$ to contain at least one quantization cell.

[6] We use the term "$\mathcal{L}$-translates of a cell $\mathcal{S}$" to denote the set $\{\mathcal{S} + \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathcal{L}\}$. The latter forms a *tessellation* if its cells cover the entire space without overlapping.

[7] It is easy to show that this result actually holds not only for a uniform distribution, but also for any sample distribution.

Since $\mathbf{F}$ is lower strictly-triangular, we have from (25) that

$$t_{j,j} = c_j - [F_{j,1} \cdots F_{j,N}]\boldsymbol{\eta}. \tag{26}$$

If $\mu_j = \mathcal{Q}(t_{j,j})$, then $t_{j,j} = v_j$, and $\eta_j = n_j$ (see (11)).

Now let us consider a vector $\mathbf{c}$ such that the target vector, at iteration $i-1$, satisfies[8]

$$t_{i-1,i-1} = \vartheta_1 - \tfrac{\Delta}{2} - \varepsilon \tag{27a}$$

$$\mathbf{t}_{i-1}^{i} = \boldsymbol{\vartheta}^1 + \mathbf{f}_i^i(\vartheta_1 - t_{i-1,i-1}), \tag{27b}$$

for some $\boldsymbol{\vartheta} \in \mathbb{U}^{N-i+1}$ and some $\varepsilon \in (0, \Delta)$. With the above target, an SF quantizer would choose $\mu_{i-1} = \mathcal{Q}(t_{i-1,i-1}) = \vartheta_1 - \Delta$, and thus $\eta_{j-1} = \varepsilon - \frac{\Delta}{2}$. Then, from (23), the cost-to-go for the SF quantizer after $i-2$ iterations can be split as

$$\|\boldsymbol{\mu}^{i-1} - \mathbf{t}_{i-1}^{i-1}\|^2_{\mathbf{G}^{i-1}} = K_{i-1}(\varepsilon - \tfrac{\Delta}{2})^2 + \|\boldsymbol{\mu}^i - \boldsymbol{\vartheta}^1 - \mathbf{f}_i^i\Delta\|^2_{\mathbf{G}^i},$$

where Lemma 2 and (27) have been used. On the other hand, from Lemma 2 and (27), the cost-to-go for the choice $\boldsymbol{\mu}^{i-1} = \boldsymbol{\vartheta}$ is $\|\boldsymbol{\vartheta}^{i-1} - \mathbf{t}_{i-1}^{i-1}\|^2_{\mathbf{G}^{i-1}} = K_i(\vartheta_1 - t_1)^2 = K_i(\frac{\Delta}{2} + \varepsilon)^2$. Therefore, the minimum difference between the cost-to-go achievable by SFQ and that of the choice $\boldsymbol{\mu}^{i-1} = \boldsymbol{\vartheta}$ is

$$\begin{aligned}\min_{\boldsymbol{\mu}^i \in \mathbb{U}^{N-i}} &\|\boldsymbol{\mu}^i - \boldsymbol{\vartheta}^1 - \mathbf{f}_i^i\Delta\|^2_{\mathbf{G}^i} - K_{i-1}\varepsilon\Delta \\ &= \min_{\boldsymbol{\mu}^i \in \mathbb{U}^{N-i}} \|\boldsymbol{\mu}^i - \mathbf{f}_i^i\Delta\|^2_{\mathbf{G}^i} - K_{i-1}\varepsilon\Delta. \end{aligned} \tag{28}$$

If (15) is not satisfied for some $i \in \{1, 2, \ldots, N\}$, then $\mathbf{f}_i^i\Delta \notin \mathbb{U}^{N-i}$, and $\nexists \boldsymbol{\mu}^i \in \mathbb{U}^{N-i}$ such that $(\boldsymbol{\mu}^i + \mathbf{f}_i^i\Delta) \in \mathcal{N}(\mathbf{G}^i)$. As a consequence, the first term on the right hand side of (28) is strictly positive. It then follows that $D(\mathbf{c}, \mathbf{u})$ is strictly larger than $D^\star(\mathbf{c})$, for sufficiently small values of ε, completing the proof. □

7. ANALYSIS OF THE RESULT

Lattice Quantization Interpretation It has been shown in the proof of Theorem 1 that (ii) is a sufficient condition for $\mathcal{L}$ to be rectangular and have a hyper-rectangular Voronoi cell. It is important to note that this can happen for a non-diagonal reconstruction Gram matrix, i.e., reconstruction vectors that are non-orthogonal, and even linearly dependent, since $\mathbf{G}$ is not required to be non-singular. It is also important to note that the converse does not necessarily hold, that is, a rectangular $\mathcal{L}$ does not ensure that condition (ii) is satisfied. More precisely, the fact that a lattice $\mathcal{L} = \mathbf{H}\mathbb{Z}^N$ is rectangular implies the existence of an integral matrix $\mathbf{M}$ with $\det\{\mathbf{M}\} = \pm 1$ such that $\mathbf{HM}$ is orthogonal. It doesn't guarantee $\mathbf{M}$ to be also lower unit-triangular, as required by (ii). On the other hand, (i) alone implies $\mathbf{HS}$ is orthogonal, and thus $\mathcal{C}(\mathbf{0}_N) = \mathbf{HS}\mathcal{Y}$ is hyper-rectangular. For a uniformly distributed $\mathbf{c}$, the MSE gap between such an SF quantizer and a lattice vector-quantizer is given by the difference between the second moments of $\mathcal{C}(\mathbf{0}_N)$ and $\mathcal{V}_0$. Although no closed from expressions are known for the second moment of $\mathcal{V}_0$ of arbitrary lattices, preliminary results suggest that it is possible to derive lower bounds for this gap from the non-integer part of the vectors $\mathbf{m}_i$ defined in (14b).

[8] Such a vector always exists, since, from (26), $t_{i-1,i-1}$ depends only on the elements $\{c_k\}_{k=1}^{i-1}$, while, from (25), $\mathbf{t}_{i-1}^i$ can be chosen independently by choosing $\{c_k\}_{k=i+1}^N$.

Reconstruction by a Single LTI Filter By letting $N \to \infty$ (and considering the *distortion per sample* $D(\mathbf{c}, \boldsymbol{\mu})/N$ as the cost function), our results can be applied to cases where reconstruction is achieved using a discrete-time LTI filter, say $R(z)$. Without loss of generality, we assume that $\lim_{z\to\infty} R(z) = 1$. In this case, the reconstruction frame vectors take the form $\psi_k = [\mathbf{0}_{k-1}^T\, r(0)\, r(1) \cdots]^T$, where $r(\cdot)$ is the impulse response of $R(z)$. This setup turns $\mathbf{F}$ and $\mathbf{H}$ into infinite dimensional Toeplitz matrices, the first column of $\mathbf{H}$ being ψ_1. In turn, $\mathbf{f}_1$ can be seen as the impulse response of a filter $F(z)$. It then follows that the orthogonality of the columns of $\mathbf{HS}$ stemming from (i) is equivalent to having $1 - F(z) = R(z)^{-1}$. This corresponds to a whitening noise-shaping quantizer, which yields minimum MSE, in the alternative white quantization noise paradigm [2]. Similarly, an SFQ satisfying (i) also minimizes the MSE, see Remark 1. On the other hand, for this case, all the vectors $\mathbf{m}_i$ (see (14b)) are equal to the impulse response (first sample removed) of $F(z)$. Thus, (ii) translates into having the impulse response of $1 - R(z)^{-1}$ being integer-valued. Hence, the standard L-th order multi-bit $\Sigma\Delta$ converter is optimal for $R(z) = (1 - z^{-1})^{-L}$. This extends the results reported in [6], obtained for $\mathbb{U} = \{-1, 1\}$.

8. CONCLUSIONS

We derived necessary and sufficient conditions that make scalar feedback quantization deterministically optimal, in the sense of generating, for any input, the quantized sequence that minimizes the 2-norm of the reconstruction error. The first condition, which depends only on the design of the scalar feedback quantizer, happens to characterize the best quantizer of this class, when a stochastic framework is adopted. The second condition depends only on the Gram matrix of the reconstruction frame, and can be satisfied for non-orthogonal, and even linearly dependent, reconstruction vectors.

9. REFERENCES

[1] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer Academic, 1992.

[2] N. Jayant and P. Noll, *Digital coding of waveforms. Principles and approaches to speech and video.* Englewood Cliffs, NJ: Prentice Hall, 1984.

[3] P. T. Boufounos and A. V. Oppenheim, "Quantization noise shaping on arbitrary frame expansions," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–12, 2006.

[4] J. J. Benedetto, A. M. Powell, and Ö. Yilmaz, "Second-order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions." *Appl. Comp. Harm. Analysis*, vol. 20, pp. 126–148, 2006.

[5] D. E. Quevedo, C. Müller, and C. G. Goodwin, "Conditions for optimality of Naïve quantized finite horizon control," *Int. J. Contr.*, vol. 80, no. 5, pp. 706–720, May 2006.

[6] A. K. Gupta and O. M. Collins, "Viterbi deconding and $\Sigma\Delta$ modulation," *IEEE Int. Symp. Information Theory*, p. 292, 2002.

[7] O. Christensen, *An introduction to frames and Riesz bases*. Boston, MA: Birkhäuser, 2003.

[8] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, UK: Cambridge University Press, 1985.

[9] J. D. Gibson and K. Sayood, "Lattice quantization," *Advances in Electronics and Electron Physics*, vol. 72, pp. 259–330, 1988.