Frame permutation quantization

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1109/CISS.2010.5464814">http://dx.doi.org/10.1109/CISS.2010.5464814</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Fri Dec 28 07:11:38 EST 2018</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/71834">http://hdl.handle.net/1721.1/71834</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher’s policy and may be subject to US copyright law. Please refer to the publisher’s site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
Frame Permutation Quantization

Ha Q. Nguyen, Vivek K Goyal, and Lav R. Varshney
Massachusetts Institute of Technology

Abstract—Frame permutation quantization (FPQ) is a new vector quantization technique using finite frames. In FPQ, a vector is encoded using a permutation source code to quantize its frame expansion. This means that the encoding is a partial ordering of the frame expansion coefficients. Compared to ordinary permutation source coding, FPQ produces a greater number of possible quantization rates and a higher maximum rate. Various representations for the partitions induced by FPQ are presented and reconstruction algorithms based on linear programming and quadratic programming are derived. Reconstruction using the canonical dual frame is also studied, and several results relate properties of the analysis frame to whether linear reconstruction techniques provide consistent reconstructions. Simulations for Gaussian sources show performance improvements over entropy-constrained scalar quantization for certain combinations of vector dimension and coding rate.

I. INTRODUCTION

Redundant representations obtained with frames are playing an ever-expanding role in signal processing due to design flexibility and other desirable properties. One such favorable property is robustness to additive noise. This robustness, carried over to quantization noise, explains the success of both ordinary oversampled analog-to-digital conversion (ADC) and Σ–Δ ADC with the canonical linear reconstruction. But the combination of frame expansions with scalar quantization is considerably more interesting and intricate because boundedness of quantization noise can be exploited in reconstruction [1]–[8] and frames and quantizers can be designed jointly to obtain favorable performance.

This paper introduces a new use of finite frames in vector quantization: frame permutation quantization (FPQ). In FPQ, permutation source coding (PSC) [9], [10] is applied to a frame expansion of a vector. This means that the vector is represented by a partial ordering of the frame coefficients (Variant I) or by signs of the frame coefficients that are larger than some threshold along with a partial ordering of the absolute values of the significant coefficients (Variant II). FPQ provides a space partitioning that can be combined with additional constraints or prior knowledge to generate a variety of vector quantizers. A simulation-based investigation shows that FPQ outperforms PSC for certain combinations of signal dimensions and coding rates. In particular, improving upon PSC at low rates provides quantizers that perform better than entropy-coded scalar quantization (ECSQ) in certain cases [11].

Beyond the exposition of the basic ideas in FPQ, the focus of this paper is on how—in analogy to works cited above—there are several decoding procedures that can sensibly be used with the encoding of FPQ. One is to use the ordinary decoding in PSC for the frame coefficients followed by linear synthesis with the canonical dual; from the perspective of frame theory, this is the natural way to reconstruct. Taking a geometric approach based on consistency yields instead optimization-based algorithms. We develop both views and find conditions on the frame used in FPQ that relate to whether the canonical reconstruction is consistent. Along with inspiring new questions about finite frames and being of interest for compression when a signal has already been acquired, FPQ may have impact on sensor design. Sensors that operate at low power and high speed by outputting orderings of signal levels rather than absolute levels have been demonstrated and are a subject of renewed interest [12], [13].

This paper is a condensed version of [14]. Among the excisions are several references, proofs, developments for a uniform source, and all further mentions of Variant II permutation codes and variable-rate coding. Preliminary results on FPQ were mentioned briefly in [15].

II. BACKGROUND

We assume fixed-rate coding and the conventional squared-error fidelity criterion \( \|x - \hat{x}\|^2 \) between source \( x \) and reproduction \( \hat{x} \). Some statements assume a known source distribution over which performance is measured in expectation.

A. Vector Quantization

A vector quantizer is a mapping from an input \( x \in \mathbb{R}^N \) to a codeword \( \hat{x} \) from a finite codebook \( \mathcal{C} \). Without loss of generality, a vector quantizer can be seen as a composition of an encoder \( \alpha : \mathbb{R}^N \to \mathcal{I} \) and a decoder \( \beta : \mathcal{I} \to \mathbb{R}^N \), where \( \mathcal{I} \) is a finite index set. The encoder partitions \( \mathbb{R}^N \) into \( |\mathcal{I}| \) regions or cells \( \{\alpha^{-1}(i)\}_{i \in \mathcal{I}} \), and the decoder assigns a reproduction value to each cell. Having \( R \) bits per component means \( |\mathcal{I}| = 2^{NR} \).

For any codebook, the encoder \( \alpha \) that minimizes \( \|x - \hat{x}\|^2 \) maps \( x \) to the nearest element of the codebook. The partition is thus composed of convex cells. Since the cells are convex, reproduction values are optimally within the corresponding cells—whether to minimize mean-squared error distortion, maximum squared error, or some other reasonable function of squared error. Reproduction values being within corresponding cells is formalized as consistency:

Definition 2.1: The reconstruction \( \hat{x} = \beta(\alpha(x)) \) is called a consistent reconstruction of \( x \) when \( \alpha(x) = \alpha(\hat{x}) \) (or equivalently \( \beta(\alpha(x)) = \hat{x} \)). The decoder \( \beta \) is called consistent when \( \beta(\alpha(x)) \) is a consistent reconstruction of \( x \) for all \( x \).
B. Permutation Source Codes

A permutation source code is a vector quantizer in which codewords are related through permutations. Permutation codes were originally introduced as channel codes by Slepian [16]. They were then applied to a specific source coding problem, through the duality between source encoding and channel decoding, by Dunn [9] and developed in greater generality by Berger et al. [10], [17], [18].

Let $\mu_1 > \mu_2 > \cdots > \mu_K$ be real numbers, and let $n_1, n_2, \ldots, n_K$ be positive integers that sum to $N$ (an ordered integer partition of $N$). The initial codeword of the codebook $C$ has the form

$$\hat{x}_{\text{init}} = (\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2, \ldots, \mu_K, \ldots, \mu_K),$$

where each $\mu_i$ appears $n_i$ times. When $\hat{x}_{\text{init}}$ has this form, we call it compatible with $(n_1, n_2, \ldots, n_K)$. The codebook is the set of all distinct permutations of $\hat{x}_{\text{init}}$. The number of codewords in $C$ is thus given by the multinomial coefficient

$$L = \frac{N!}{n_1! n_2! \cdots n_K!}.$$  

The permutation structure of the codebook enables low-complexity nearest-neighbor encoding [10]: map $x$ to the codeword $\hat{x}$ whose components have the same order as $x$; in other words, replace the $n_1$ largest components of $x$ with $\mu_1$, the $n_2$ next-largest components of $x$ with $\mu_2$, and so on. Since the complexity of sorting a vector of length $N$ is $O(N \log N)$ operations, the encoding complexity for PSC is much lower than with an unstructured source code and only $O(\log N)$ times higher than scalar quantization.

The per-component rate is defined as

$$R = N^{-1} \log_2 L.$$  

Under certain symmetry conditions on the source distribution, all codewords are equally likely so the rate cannot be reduced by entropy coding.

**Partition Properties:** The partition induced by a PSC is completely determined by the integer partition $(n_1, n_2, \ldots, n_K)$. Specifically, the encoding mapping can index the permutation $P$ that places the $n_1$ largest components of $x$ in the first $n_1$ positions (without changing the order within those $n_1$ components), the $n_2$ next-largest components of $x$ in the next $n_2$ positions, and so on; the $\mu_i$s are actually immaterial. This encoding is placing all source vectors $x$ such that $P^{-1}x$ is $n$-descending in the same partition cell, defined as follows.

**Definition 2.2:** Given an ordered integer partition $n = (n_1, n_2, \ldots, n_K)$ of $N$, a vector in $\mathbb{R}^N$ is called $n$-descending if its $n_1$ largest entries are in the first $n_1$ positions, its $n_2$ next-largest components are in the next $n_2$ positions, etc.

The property of being $n$-descending is to be descending up to the arbitrariness specified by the integer partition $n$.

Because this is nearest-neighbor encoding for some codebook, the partition cells must be convex. Furthermore, multiplying $x$ by any nonnegative scalar does not affect the encoding, so the cells are convex cones. We develop a convenient representation for the partition in Section III.

**Codebook Optimization:** Assume that $x$ is random and that the components of $x$ are i.i.d. Let $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_N$ denote the order statistics of random vector $x = (x_1, \ldots, x_N)$ and $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_N$ denote the order statistics of random vector $x' = (x_1', \ldots, x_N')$.\footnote{For consistency with earlier literature on PSCs, we are reversing the usual sorting of order statistics [19].} For a given initial codeword $\hat{x}_{\text{init}}$, the per-letter distortion of an optimally-encoded PSC is given by

$$D = N^{-1} E \left[ \sum_{i=1}^{K} \sum_{t \in I_i} (\xi_t - \mu_i)^2 \right]$$

where $I_i$s are the sets of indexes generated by the integer partition:

$$I_1 = \{1, 2, \ldots, n_1\},$$

$$I_2 = \{n_1 + 1, n_1 + 2, \ldots, n_1 + n_2\},$$

etc. These distortions can be deduced simply by examining which components of $x$ are mapped to which elements of $\hat{x}_{\text{init}}$.

Optimization of (4) over both $\{n_i\}_{i=1}^{K}$ and $\{\mu_i\}_{i=1}^{K}$ subject to (3) is difficult, partly due to the integer constraint of the partition. However, given an integer partition $(n_1, n_2, \ldots, n_K)$, the optimal initial codeword can be determined easily from the means of the order statistics:

$$\mu_i = n_i^{-1} \sum_{t \in I_i} E[x_t].$$

The analysis of [17] shows that when $N$ is large, the optimal partition gives performance equal to ECSQ of $x$. Performance does not strictly improve with increasing $N$; permutation codes outperform ECSQ for certain combinations of block size and rate [11].

C. Frame Definitions and Classifications

In this paper, we use frame expansions only for quantization using PSCs, which rely on order relations of real numbers. Therefore we limit ourselves to real finite frames.

**Definition 2.3:** A set of $N$-dimensional vectors, $\Phi = \{\phi_k\}_{k=1}^{M} \subset \mathbb{R}^N$, is called a frame if there exist a lower frame bound, $A > 0$, and an upper frame bound, $B < \infty$, such that, for all $x \in \mathbb{R}^N$,

$$A||x||^2 \leq \sum_{k=1}^{M} |\langle x, \phi_k \rangle|^2 \leq B||x||^2.$$  

\hspace{1cm} (7a)

The matrix $F \in \mathbb{R}^{M \times N}$ with $k$th row equal to $\phi_k$ is called the analysis frame operator. Equivalent to (7a) in matrix form is

$$A|x|^2 \leq \sum_{k=1}^{M} |\langle x, \phi_k \rangle|^2 \leq B||x||^2.$$  

\hspace{1cm} (7b)

where $I_N$ is the $N \times N$ identity matrix.

A frame is called a tight frame if the frame bounds can be chosen to be equal. A frame is an equal-norm frame if all of its vectors have the same norm. If an equal-norm frame is normalized to have all vectors of unit norm, we call it a unit-norm frame. A unit-norm tight frame (UNTF) must satisfy $F^*F = (M/N)I_N$. \hspace{1cm} (7b)
Definition 2.4: The real harmonic tight frame (HTF) of \( M \) vectors in \( \mathbb{R}^N \) is defined for even \( N \) by
\[
\phi_{k+1} = \sqrt{\frac{2}{N}} \left[ \begin{array}{c} \cos \frac{k\pi}{M} \cos \frac{3k\pi}{M} \cos (N-1)k\pi \vspace{1em} \\ \sin \frac{k\pi}{M} \sin \frac{3k\pi}{M} \sin (N-1)k\pi \end{array} \right],
\]
and for odd \( N \) by
\[
\phi_{k+1}^* = \sqrt{\frac{2}{N}} \left[ \begin{array}{c} \cos \frac{2k\pi}{M} \cos \frac{4k\pi}{M} \cos (N-1)k\pi \vspace{1em} \\ \sin \frac{2k\pi}{M} \sin \frac{4k\pi}{M} \sin (N-1)k\pi \end{array} \right],
\]
where \( k = 0, 1, \ldots, M - 1 \). The modulated harmonic tight frames are defined by
\[
\psi_k = \gamma(-1)^k \phi_k, \quad \text{for } k = 1, 2, \ldots, M,
\]
where \( \gamma = 1 \) or \( \gamma = -1 \) (fixed for all \( k \)).

Classification of frames is often up to some unitary equivalence. Holmes and Paulsen [20] proposed several types of equivalence relations between frames. In particular, for two frames in \( \mathbb{R}^N \), \( \Phi = \{ \phi_k \}_{k=1}^M \) and \( \Psi = \{ \psi_k \}_{k=1}^M \), we say \( \Phi \) and \( \Psi \) are

(i) type I equivalent if there is an orthogonal matrix \( U \) such that \( \psi_k = U \phi_k \) for all \( k \);
(ii) type II equivalent if there is a permutation \( \sigma(\cdot) \) on \( \{1, 2, \ldots, M\} \) such that \( \psi_k = \phi_{\sigma(k)} \) for all \( k \); and
(iii) type III equivalent if there is a sign function in \( k \), \( \delta(k) = \pm 1 \) such that \( \psi_k = \delta(k) \phi_k \) for all \( k \).

Another important subclass of UNTFs is defined as follows:

Definition 2.5 [21], [22]): A UNTF \( \Phi = \{ \phi_k \}_{k=1}^M \subset \mathbb{R}^N \) is called an equiangular tight frame (ETF) if there exists a constant \( a \) such that \( \langle \phi_k, \phi_j \rangle = a \) for all \( 1 \leq k < j \leq M \). ETFs are sometimes called optimal Grassmannian frames or 2-uniform frames.

In our analysis of FPQ, we will find that restricted ETFs—where the absolute value constraint can be removed from Definition 2.5—play a special role. In matrix view, a restricted ETF satisfies \( F^* F = (M/N)I_N \) and \( FF^* = (1-a)I_M + aJ_M \), where \( J_M \) is the all-1s matrix of size \( M \times M \). The following proposition specifies the restricted ETFs for the codimension-1 case.

Proposition 2.6: For \( M = N+1 \), the family of all restricted ETFs is constituted by the Type I and Type II equivalents of modulated HTFs.

Finally, the following property of modulated HTFs in the \( M = N+1 \) case will be very useful.

Proposition 2.7: If \( M = N+1 \) then a modulated harmonic tight frame is a zero-sum frame, i.e., each column of the analysis frame operator \( F \) sums to zero.

D. Reconstruction from Frame Expansions

A central use of frames is to formalize the reconstruction of \( x \in \mathbb{R}^N \) from the frame expansion \( y_k = \langle x, \phi_k \rangle \), \( k = 1, 2, \ldots, M \), or estimation of \( x \) from degraded versions of the frame expansion. Using the analysis frame operator we have \( y = Fx \), and (7) implies the existence of at least one linear synthesis operator \( G \) such that \( GF = I_N \). A frame with analysis frame operator \( G^* \) is then said to be dual to \( \Phi \).

The frame condition (7) also implies that \( F^* F \) is invertible, so the Moore–Penrose inverse (pseudo-inverse) of the frame operator \( F^* = (F^* F)^{-1}F^* \) exists and is a valid synthesis operator. Using the pseudo-inverse for reconstruction has several important properties including an optimality for mean-squared error (MSE) under assumptions of uncorrelated zero-mean additive noise and linear synthesis. Reconstruction using \( F^* \) is called canonical reconstruction and the corresponding frame is called the canonical dual. In this paper, we use the term linear reconstruction for reconstruction using an arbitrary linear operator.

When \( y \) is quantized to \( \hat{y} \), it is possible for the quantization noise \( \hat{y} - y \) to have mean zero and uncorrelated components; this occurs with subtractive dithered quantization [23] or under certain asymptotics [24]. In this case, the optimality of canonical reconstruction holds. However, it should be noted that even with these restrictions, canonical reconstruction is optimal only among linear reconstructions.

When nonlinear construction is allowed, quantization noise may behave fundamentally differently than other additive noise. The key is that a quantized value gives hard constraints that can be exploited in reconstruction. For example, suppose that \( \hat{y} \) is obtained from \( y \) by rounding each element to the nearest multiple of a quantization step size \( \Delta \). Then knowledge of \( y_m \) is equivalent to knowing
\[
y_k \in \left[ \hat{y}_k - \frac{1}{2}\Delta, \hat{y}_k + \frac{1}{2}\Delta \right].
\]
Geometrically, \( \langle x, \phi_k \rangle = \hat{Y}_k - \frac{1}{2}\Delta \) and \( \langle x, \phi_k \rangle = \hat{y}_k + \frac{1}{2}\Delta \) are hyperplanes perpendicular to \( \phi_k \), and (10) expresses that \( x \) must lie between these hyperplanes. Using the upper and lower bounds on all \( M \) components of \( y \), the constraints on \( x \) imposed by \( \hat{y} \) are readily expressed as [2]
\[
\left[ \begin{array}{c} F \vspace{1em} \\ -F \end{array} \right] x \leq \left[ \begin{array}{c} \frac{1}{2}\Delta + \hat{y} \vspace{1em} \\ \frac{1}{2}\Delta - \hat{y} \end{array} \right],
\]
where the inequalities are elementwise.

III. FRAME PERMUTATION QUANTIZATION

FPQ is simply PSC applied to a frame expansion.

A. Encoder Definition and Canonical Decoding

Definition 3.1: A frame permutation quantizer with analysis frame \( F \in \mathbb{R}^{M \times N} \), integer partition \( m = (m_1, m_2, \ldots, m_K) \), and initial codeword \( \hat{y}_{init} \) compatible with \( m \) encodes \( x \in \mathbb{R}^N \) by applying a permutation source code with integer partition \( m \) and initial codeword \( \hat{y}_{init} \) to \( Fx \).

The canonical decoding gives \( \hat{x} = F^* \hat{y} \), where \( \hat{y} \) is the PSC
reconstruction of \( y \). We sometimes use the triple \((F, m, \hat{y}_{\text{init}})\) to refer to such an FPQ.

The result of the encoding can be expressed as a permutation \( P \) from the permutation matrices of size \( M \). The permutation is such that \( PFx \) is \( m \)-descending. For uniqueness in the representation \( P \) chosen from the set of permutation matrices, we can specify that the first \( m_1 \) components of \( Py \) are kept in the same order as they appeared in \( y \), the next \( m_2 \) components of \( Py \) are kept in the same order as they appeared in \( y \), etc. Then \( P \) is in a subset \( \mathcal{G}(m) \) of the \( M \times M \) permutation matrices and

\[
|\mathcal{G}(m)| = \frac{M!}{m_1! m_2! \cdots m_K!}.
\]

Notice that the encoding uses the integer partition \( m \) but not the initial codeword \( \hat{y}_{\text{init}} \). The PSC reconstruction of \( y \) is \( P^{-1} \hat{y}_{\text{init}} \), so the canonical decoding gives \( F^\dagger P^{-1} \hat{y}_{\text{init}} \).

The size of the set \( \mathcal{G}(m) \) is analogous to the codebook size in (2), and the per-component rate of FPQ is thus defined as

\[
R = N^{-1} \log_2 \frac{M!}{m_1! m_2! \cdots m_K!}.
\]

### B. Expressing Consistency Constraints

Suppose FPQ encoding of \( x \in \mathbb{R}^N \) with frame \( F \in \mathbb{R}^{M \times N} \), integer partition \( m = (m_1, m_2, \ldots, m_K) \), and initial codeword \( \hat{y}_{\text{init}} \) compatible with \( m \) results in permutation \( P \in \mathcal{G}(m) \). We would like to express constraints on \( x \) that are specified by \((F, m, \hat{y}_{\text{init}}, P)\). This will provide an explanation of the partitions induced by FPQ and lead to reconstruction algorithms in Section III-C.

Knowing that a vector is \( m \)-descending is a specification of many inequalities. Recall the definitions of the index sets generated by an integer partition given in (5), and use the same notation with \( n_k \)'s replaced by \( m_k \)'s. Then \( z \) being \( m \)-descending implies that for any \( i < j \),

\[
z_k \geq z_\ell \quad \text{for every } k \in \mathcal{I}_i \text{ and } \ell \in \mathcal{I}_j.
\]

By transitivity, considering every \( i < j \) gives redundant inequalities. Taking only \( j = i+1 \), we obtain a full description

\[
z_k \geq z_\ell \quad \text{for every } k \in \mathcal{I}_i \text{, } \ell \in \mathcal{I}_{i+1} \text{ with } i = 1, \ldots, K - 1.
\]

For one fixed \((i, \ell)\) pair, (14) gives \(|\mathcal{I}_i| = m_i \) inequalities, one for each \( k \in \mathcal{I}_i \). These inequalities can be gathered into an elementwise matrix inequality as

\[
\begin{bmatrix}
0_{m_i \times M_i - 1} & I_{m_i} & 0_{m_i \times (M - M_i)} \\
0_{m_i \times (M - M_i - 1)} & 1_{m_i \times 1} & 0_{m_i \times (M - \ell)}
\end{bmatrix} z
\geq
\begin{bmatrix}
0_{m_i \times (\ell - 1)} & I_{m_i} & 0_{m_i \times (M - M_i - \ell)} \\
0_{m_i \times (M - \ell - 1)} & -1_{m_i \times 1} & 0_{m_i \times (M - \ell)}
\end{bmatrix} z
\]

where \( M_k = m_1 + m_2 + \cdots + m_k \), or \( D_i^{(m)} z \geq 0_{m_i \times 1} \) where

\[
D_i^{(m)} = \begin{bmatrix}
0 & I & 0 & \cdots & 0 \\
0 & 0 & I & \cdots & 0 \\
& \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & I & 0 \\
0 & \cdots & 0 & 0 & I
\end{bmatrix}
\]

is an \( m_i \times M \) differencing matrix. Allowing \( \ell \) to vary across \( \mathcal{I}_{i+1} \), we define the \( m_i m_{i+1} \times M \) matrix

\[
D_i^{(m)} = \begin{bmatrix}
D_i^{(m)}_{1, M_i + 1} \\
D_i^{(m)}_{1, M_i + 2} \\
\vdots \\
D_i^{(m)}_{1, M_i + m_{i+1}}
\end{bmatrix}
\]

and express all of (14) for one fixed \( i \) as \( D_i^{(m)} z \geq 0_{m_i m_{i+1} \times 1} \).

Continuing our recursion, it only remains to gather the inequalities (14) across \( i = 1, 2, \ldots, K - 1 \). Let

\[
D^{(m)} = \begin{bmatrix}
D_1^{(m)} \\
D_2^{(m)} \\
\vdots \\
D_{K-1}^{(m)}
\end{bmatrix}
\]

which has

\[
L(m) = \sum_{i=1}^{K-1} m_i m_{i+1}
\]

rows. The property of \( z \) being \( m \)-descending can be expressed as \( D^{(m)} z \geq 0_{L(m) \times M} \).

Now we can apply these representations to FPQ. Since \( PFx \) is \( m \)-descending, consistency is simply expressed as

\[
D^{(m)} PFx \geq 0.
\]

### C. Consistent Reconstruction Algorithms

The constraints (17) specify unbounded sets. To be able to decode FPQs in analogy to [2, Table I], we require some additional constraints. Two examples are developed in [14]: a source \( x \) bounded to \([-1/2, 1/2]^N \) (e.g., having an i.i.d. uniform distribution over \([-1/2, 1/2]^N \)) or having an i.i.d. standard Gaussian distribution. The Gaussian case is recounted here.

Suppose \( x \) has i.i.d. Gaussian components with mean zero and unit variance. Since the source support is unbounded, something beyond consistency must be used in reconstruction. Here we use a quadratic program to find a good bounded, consistent estimate and combine this with the average value of \( \|x\| \).

The problem with using (17) combined with maximization of minimum slackness alone (without any additional boundedness constraints) is that for any purported solution, multiplying by a scalar larger than 1 will increase the slackness of all the constraints. Thus, any solution technique will naturally and correctly have \( \|\hat{x}\| \rightarrow \infty \). Actually, because the partition cells are convex cones, we should not hope to recover the radial component of \( x \) from the partition. Instead, we should only hope to recover a good estimate of \( x/\|x\| \).

To estimate the angular component \( x/\|x\| \) from the partition, it would be convenient to maximize minimum slackness while also imposing a constraint of \( \|\hat{x}\| = 1 \). Unfortunately, this is a nonconvex constraint. It can be replaced by \( \|\hat{x}\| \leq 1 \).
Because slackness is proportional to $\|\hat{x}\|$. This suggests the optimization

maximize $\delta$ subject to $\|x\| \leq 1$ and $D^{(m)} P F x \geq \delta I_{L(m)} x . 1$.

Denoting the $x$ at the optimum by $\hat{x}_{\text{ang}}$, we still need to choose the radial component, or length, of $\hat{x}$.

For the $N(0, I_N)$ source, the mean length is [25]

$$E[\|x\|] = \frac{\sqrt{2\pi}}{\beta(N/2, 1/2)} \approx \sqrt{N - 1/2}.$$ 

We can combine this with $\hat{x}_{\text{ang}}$ to obtain a reconstruction $\hat{x}$. This method is described explicitly in [14, Alg. 3].

IV. CONDITIONS ON THE CHOICE OF FRAME

In this section, we provide necessary and sufficient conditions so that a linear reconstruction is also consistent. We first consider a general linear reconstruction, $\hat{x} = R\hat{y}$, where $R$ is some $N \times M$ matrix and $\hat{y}$ is a decoding of the PSC of $y$. We then restrict attention to canonical reconstruction, where $R = F^t$. For each case, we describe all possible choices of a “good” frame $F$, in the sense of the consistency of the linear reconstruction.

A. Arbitrary Linear Reconstruction

We begin by introducing a useful term.

Definition 4.1: A matrix is called column-constant when each column of the matrix is a constant. The set of all $M \times M$ column-constant matrices is denoted $\mathcal{J}$.

We now give our main results for arbitrary linear reconstruction combined with FPQ decoding of an estimate of $y$.

Theorem 4.2: Suppose $A = FR = aI_M + J$ for some $a \geq 0$ and $J \in \mathcal{J}$. Then the linear reconstruction $\hat{x} = R\hat{y}$ is consistent with Variant I FPQ encoding using frame $F$, an arbitrary integer partition and an arbitrary Variant I initial codeword compatible with it.

The key point in the proof of Theorem 4.2 is showing that the inequality

$$D^{(m)} P A P^{-1} \hat{y}_{\text{init}} \geq 0,$$ 

where $A = FR$, holds for every integer partition $m$ and every initial codeword $\hat{y}_{\text{init}}$ compatible with it. It turns out that the form of matrix $A$ given in Theorem 4.2 is the unique form that guarantees that (18) holds for every pair $(m, \hat{y}_{\text{init}})$. In other words, the condition on $A$ that is sufficient for any integer partition $m$ and any initial codeword $\hat{y}_{\text{init}}$ compatible with it is also necessary for consistency for every pair $(m, \hat{y}_{\text{init}})$.

Theorem 4.3: Consider FPQ using frame $F$ with $M \geq 3$. If linear reconstruction $\hat{x} = R\hat{y}$ is consistent with every integer partition and every initial codeword compatible with it, then matrix $A = FR$ must be of the form $aI_M + J$, where $a \geq 0$ and $J \in \mathcal{J}$.

B. Canonical Reconstruction

We now restrict the linear reconstruction to use the canonical dual; i.e., $R$ is restricted to be the pseudo-inverse $F^t = (F^* F)^{-1} F^*$. The following corollary characterizes the non-trivial frames for which canonical reconstructions are consistent.

Corollary 4.4: Consider FPQ using frame $F$ with $M > N$ and $M \geq 3$. For canonical reconstruction to be consistent with every integer partition and every initial codeword compatible with it, it is necessary and sufficient to have $M = N + 1$ and $A = FF^t = I_M - \frac{1}{M} J_M$, where $J_M$ is the $M \times M$ all-1s matrix.

We continue to add more constraints to frame $F$. By imposing tightness and unit norm on our analysis frame, we can progress a bit further from Corollary 4.4 to derive the form of $F^*$.

Corollary 4.5: Consider FPQ using unit-norm tight frame $F$ with $M > N$ and $M \geq 3$. For canonical reconstruction to be consistent for every integer partition and every initial codeword compatible with it, it is necessary and sufficient to have $M = N + 1$ and

$$FF^* = \begin{bmatrix} 1 & -\frac{1}{N} & \cdots & -\frac{1}{N} \\ -\frac{1}{N} & 1 & \cdots & -\frac{1}{N} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{N} & -\frac{1}{N} & \cdots & 1 \end{bmatrix}.$$ 

Recall that a UNTF that satisfies (19) is a restricted ETF. Therefore Corollary 4.5 together with Proposition 2.6 gives us a complete characterization of UNTFs that are “good” in the sense of canonical reconstruction being consistent.

Corollary 4.6: Consider FPQ using unit-norm tight frame $F$ with $M > N$ and $M \geq 3$. For canonical reconstruction to be consistent for every integer partition and every initial codeword compatible with it, it is necessary and sufficient for $F$ to be a modulated HTF or a Type I or Type II equivalent.

V. SIMULATIONS

In this section, we provide simulations to demonstrate some properties of FPQ and to demonstrate that FPQ can give performance better than ECSQ and ordinary PSC for certain combinations of signal dimension and rate. For every data point shown, the distortion represents a sample mean estimate of $N^{-1} E[\|x - \hat{x}\|^2]$ over at least $10^6$ trials. Testing was done with exhaustive enumeration of the relevant integer partitions. This makes the complexity of simulation high, and thus experiments are only shown for small $N$ and $M$. Recall the encoding complexity of FPQ is low, $O(M \log M)$ operations. The decoding complexity is polynomial in $M$, and in some applications it could be worthwhile to precompute the entire codebook at the decoder. Thus much larger values of $N$ and $M$ than used here may be practical.

Gaussian source. Let $x$ have the $N(0, I_N)$ distribution. Figure 1 summarizes the performance of FPQ with decoding using the algorithm described in Section III-C. Also shown are the performance of entropy-constrained scalar quantization and
the distortion–rate bound. Of course, the distortion–rate bound can only be approached with \( N \to \infty \); it is not presented as a competitive alternative to FPQ for \( N = 4 \) and \( N = 5 \).

We have not provided an explicit comparison to ordinary PSC because, due to rotational-invariance of the Gaussian source, FPQ with any orthonormal basis as the frame is identical to PSC. (The modulated harmonic tight frame with \( M = N \) is an orthonormal basis.) PSC and FPQ are sometimes better than ECSQ; increasing \( M \) gives more operating points and a higher maximum rate; and \( M = N + 1 \) seems especially attractive.

**REFERENCES**


---

A discussion of the density of PSC rates is given in [26, App. B].

Fig. 1. Performance of Variant I FPQ on an i.i.d. \( N(0, 1) \) using modulated harmonic tight frames ranging in size from \( N = 4 \) to \( N = 7 \). Performance of PSC is not shown because it is equivalent to FPQ with \( M = N \) for this source. Also plotted are the performance of entropy-constrained scalar quantization and the distortion-rate bound.