TRANSFORM CODING

A way to use scalar quantizers to efficiently encode dependent sources.

Let $X_1, X_2, \ldots$ be a stationary or WSS random process.
Instead of quantizing each X_i with a scalar quantizer

1) Parse data sequence into blocks of some length k,

$$X_1, X_2, \ldots = \underline{X}_1, \underline{X}_2 \ldots$$

where $\underline{X}_i = (X_{(i-1)k+1}, \ldots, X_{ik})$; encode each $\underline{X}_i$ as follows.

2) Find and scalar quantize the coefficients of an orthonormal expansion of $\underline{X}$ $(= \underline{X}_i)$ with respect to some orthonormal basis $\underline{t}_1, \ldots, \underline{t}_k$. $(\underline{t}_i = (t_{i1}, \ldots, t_{ik})^t)$

a) Find coefficients $U_1, \ldots, U_k$ such that

$$\underline{X} = U_1\underline{t}_1 + \ldots + U_k\,\underline{t}_k$$

b) Scalar quantize each U_i with a quantizer that is optimized for it.

The quantized values are

$$V_1 = Q_1(U_1), \ldots, V_k = Q_k(U_k)$$

The encoder output is

$$\underline{B} = e_1(U_1), \ldots, e_k(U_k)$$

3) The decoder receives $e_1(U_1), \ldots, e_k(U_k)$ and produces

$$\underline{Y} = V_1\underline{t}_1 + \ldots + V_k\,\underline{t}_k$$

Key Ideas:

A: Choose $\underline{t}_1, \ldots, \underline{t}_k$ so that $U_1, \ldots, U_k$ are uncorrelated.

(If they were correlated, then decorrelating would seem to do better.)

B: Choose $\underline{t}_1, \ldots, \underline{t}_k$ so that relatively few U_i's are large, and use higher rate scalar quantizers on U_i's that are larger on the average, than on U_i's that are smaller on the average. (This is sometimes called "energy compaction".)

It remains to be seen that more rate is saved on the small U_i's, than is spent on the large U_i's.

We will show: A $\Leftrightarrow$ B.

MATHEMATICAL DETAILS (FROM LINEAR ALGEBRA)

Definitions:

- R^k = set of real-valued k-dimensional vectors of the form $\underline{x} = (x_1,\ldots,x_k)$.
  We often think of vectors as column vectors.
- Inner or dot product: $(\underline{x},\underline{y}) = (\underline{y},\underline{x}) = \underline{x}^t\underline{y} = \underline{y}^t\underline{x} = \sum_{i=1}^{k} x_i\, y_i$
- Length of $\underline{x}$: $||\underline{x}|| = \sqrt{\sum_{i=1}^{k} x_i^2} = \sqrt{(\underline{x},\underline{x})} = \sqrt{\underline{x}^t\underline{x}}$

  Fact: $||\underline{x} + \underline{y}||^2 = (\underline{x}+\underline{y})^t(\underline{x}+\underline{y}) = ||\underline{x}||^2 + 2(\underline{x},\underline{y}) + ||\underline{y}||^2$
- Orthogonality: $\underline{x}$ and $\underline{y}$ are orthogonal if $\underline{x}^t\underline{y} = 0$.
- Linear independence: $\underline{x}_1,\ldots,\underline{x}_m$ are linearly independent if there exist no coefficients $u_1,\ldots,u_m$ such that $u_1\underline{t}_1 + \ldots + u_m\,\underline{t}_k = \underline{0}$, except $u_1 = u_2 = \ldots = u_m = 0$.
- Basis: A basis for R^k is a linearly independ. set of vectors $\{\underline{t}_1,\ldots,\underline{t}_m\}$ in R^k that span R^k. That is, every vector $\underline{x} \in R^k$ is a linear combination of the members of the set; i.e. there are coefficients $u_1,\ldots,u_m$ such that $\underline{x} = u_1\underline{t}_1 + \ldots + u_m\,\underline{t}_k$
- Orthonormal basis for R^k: $\{\underline{t}_1,\ldots,\underline{t}_k\}$ such that that members of the basis are:

  orthogonal: $\underline{t}_i^t\,\underline{t}_j = 0$ when $i \neq j$, and normalized: $||\underline{t}_i|| = 1$ for each i

Facts:

- Every basis for R^k has exactly k elements.
- There exists an orthonormal basis for every k,
  e.g. $\{(1,0,\ldots,0), (0,1,0,\ldots,0), \ldots, (0,\ldots,0,1)\}$
- Given a basis $\{\underline{t}_1,\ldots,\underline{t}_m\}$ and a vector $\underline{x}$ there is exists one and only one set of coefficients $u_1,\ldots,u_m$ such that $\underline{x} = u_1\underline{t}_1 + \ldots\, u_m\,\underline{t}_k$
- Given an orthonormal basis $\underline{t}_1,\ldots,\underline{t}_k$ and a vector $\underline{x}$ the coefficients $u_1,\ldots,u_m$ such that $\underline{x} = u_1\underline{t}_1 + \ldots + u_m\,\underline{t}_k$ are

  $u_i = \underline{t}_i^t\,\underline{x},\ i = 1,\ldots,k$

  Proof: If $\underline{x} = u_1\underline{t}_1 + \ldots + u_k\,\underline{t}_k$, then

  $\underline{t}_i^t\,\underline{x} = \underline{t}_i^t\,(u_1\underline{t}_1 + \ldots + u_k\,\underline{t}_k) = u_1\underline{t}_i^t\,\underline{t}_1 + \ldots + u_k\,\underline{t}_i^t\,\underline{t}_k = u_i$ since $\underline{t}_i$'s orthonormal
- If $\underline{u} = (u_1,\ldots,u_k)$ is the vector of coefficients of the expansion of $\underline{x}$ with respect to orthonormal basis $\{\underline{t}_1,\ldots,\underline{t}_m\}$, then

  $||\underline{u}||^2 = ||\underline{x}||^2$

  Proof: $||\underline{x}||^2 = ||u_1\underline{t}_1 + \ldots + u_m\underline{t}_k||^2 = (u_1\underline{t}_1 + \ldots + u_m\underline{t}_k)^t(u_1\underline{t}_1 + \ldots + u_m\underline{t}_k)$

  $= u_1^2 + \ldots + u_k^2$ because $\underline{t}_i^t\,\underline{t}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

  $= ||\underline{u}||^2$

Recall Step 2a of transform coding

2) Given orthonormal basis $\underline{t}_1,\ldots,\underline{t}_k$. $(\underline{t}_i = (t_{i1},\ldots,t_{ik})^t)$

(a) Find coefficients $U_1,\ldots,U_k$ such that $\underline{X} = U_1\underline{t}_1 + \ldots + U_k\,\underline{t}_k$

Therefore,

$$U_i = \underline{t}_i^t\,\underline{X} = \sum_{j=1}^{k} t_{ij}\,X_j$$

Equivalently, we can compute the coefficients via a matrix multiply

$$\underline{U} = (U_1,\ldots,U_k)\quad T\underline{X}$$

where T is $k\times k$ matrix $T = \begin{bmatrix} --\underline{t}_1^t-- \\ --\underline{t}_2^t-- \\ \ldots \\ --\underline{t}_k^t-- \end{bmatrix} = \begin{bmatrix} t_{11}\ t_{12}\ \ldots\ t_{1k} \\ t_{21}\ t_{22}\ \ldots\ t_{2k} \\ \ldots \\ t_{k1}\ t_{k2}\ \ldots\ t_{kk} \end{bmatrix}$

More Definitions:

- In this context the matrix T is called a *transform*, and $T\underline{X}$ is called a transformation of $\underline{X}$. $\underline{U}$ is the vector of *transform coefficients*.
- A matrix with orthonormal rows, such as the ones we are interested in, is called "orthogonal". ("Othonormal" would be a better name, but "orthogonal" is what the mathematicians use.)

Facts: The following are equivalent statements

(a) T is an orthogonal matrix, (b) Rows of T are orthormal

(c) Columns of T are orthonormal, (d) $||T\underline{x}|| = ||\underline{x}||$ for all $\underline{x}$

(e) $T^{-1} = T^t$, (f) T^{-1} is orthogonal.

Proofs:

(a) $\Leftrightarrow$ (b): By definition of an orthogonal matrix

(b) $\Leftrightarrow$ (e): Notice that $TT^t = \begin{bmatrix} --\underline{t}_1^t-- \\ --\underline{t}_2^t-- \\ \ldots \\ --\underline{t}_k^t-- \end{bmatrix} \begin{bmatrix} | & | & & | \\ \underline{t}_1 & \underline{t}_2 & \cdots & \underline{t}_k \\ | & | & & | \end{bmatrix}$ equals $I = \begin{bmatrix} 1\ 0\ \ldots..\ 0 \\ 0\ 1\ 0\ ..\ 0 \\ \\ 0\ \ldots..\ 0\ 1 \end{bmatrix}$

if and only if the rows of T are orthormal. Thus T^t is the inverse of T if and only if its rows are orthormal. (Note that by definition of the inverse $T^{-1}T = I = TT^{-1}$.)

(c) $\Leftrightarrow$ (e): By a similar argument $T^tT = I$ iff the columns of T are orthormal.

(b) $\Rightarrow$ (d): Proved earlier.

(e) $\Rightarrow$ (d): (e) $\Rightarrow$ $||T\underline{x}||^2 = (T\underline{x},T\underline{x}) = (T\underline{x})^tT\underline{x} = \underline{x}^tT^tT\underline{x} = \underline{x}^t\underline{x} = ||\underline{x}||^2$

(d) $\Rightarrow$ (e): Assuming $||T\underline{x}|| = ||\underline{x}||$ for all $\underline{x}$, we have

$0 = ||T\underline{x}|| - ||\underline{x}|| = (T\underline{x},T\underline{x}) - (\underline{x},\underline{x}) = (T\underline{x})^tT\underline{x} - \underline{x}^t\underline{x} = \underline{x}^tT^tT\underline{x} - \underline{x}^tI\underline{x} = \underline{x}^t(T^tT-I)\underline{x}$

since this holds for all $\underline{x}$ it must be that $T^tT=I$, i.e. $T^t = T^{-1}$.

(e) $\Leftrightarrow$ (f): Elementary.

Fact: For an orthogonal transformation $||T\underline{x}-T\underline{y}|| = ||\underline{x}-\underline{y}||$

Proof: $||T\underline{x}-T\underline{y}||^2 = (T\underline{x}-T\underline{y},T\underline{x}-T\underline{y}) = (T\underline{x}-T\underline{y})^t(T\underline{x}-T\underline{y})$

$= (T(\underline{x}-\underline{y}))^tT(\underline{x}-\underline{y}) = (\underline{x}-\underline{y})^tT^tT(\underline{x}-\underline{y})$

$= (\underline{x}-\underline{y})^t(\underline{x}-\underline{y}) = ||\underline{x}-\underline{y}||^2$

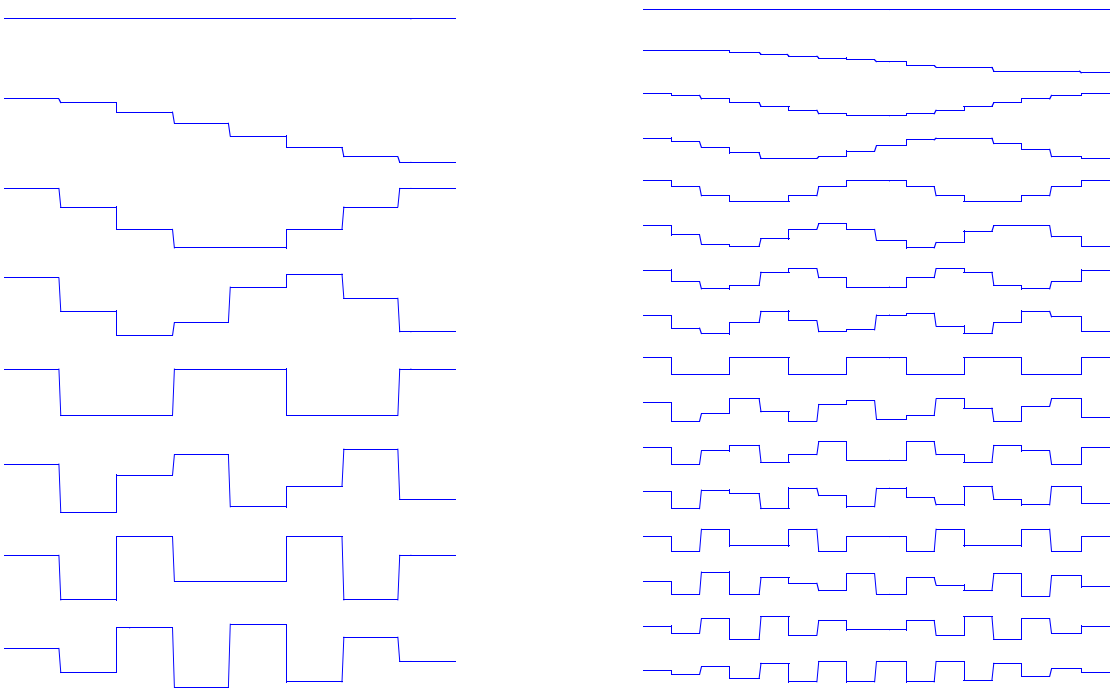
Because an orthogonal transformation preserves the length of vectors and also the distance between vectors, we may view it as essentially a rotation. (There might also be axis flips.

Examples of orthogonal transformations:

- DFT discrete Fourier transform
- DCT discrete cosine transform
- DWT discrete wavelet transform
- WHT Walsh-Hadamard
- Two-dimensional versions of the above
- KLT Karhunen-Loeve transform

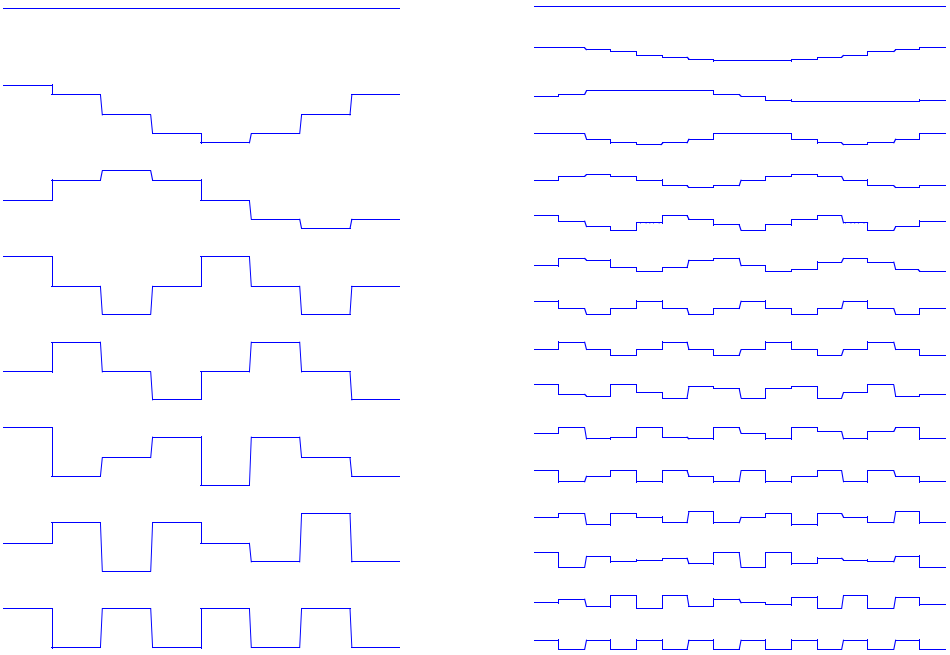
EXAMPLE: DCT BASIS VECTORS (ONE-DIMENSIONAL)

k = 8 **k = 16**

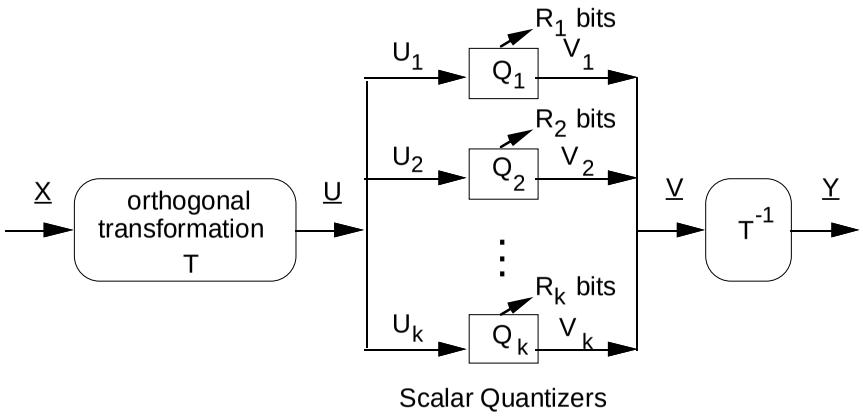


Example: (Real) FFT Basis Vectors (one-dimensional)

k = 8 **k = 16**



TRANSFORM CODING



Scalar Quantizers

Performance:

Rate: $R = \frac{1}{k} \sum_{i=1}^{k} R_i$, where R_i is rate of Q_i (SQ-FL or SQ-VL).

The transform code is *fixed rate* if its scalar quantizers are fixed-rate (i.e. SQ-FL) and *variable rate*, if its scalar quantizers are variable-rate (i.e. SQ-VL).

Key Property: $\|\underline{X}-\underline{Y}\| = \|\underline{U}-\underline{V}\|$

$\Rightarrow$ Distortion is

$$D = \frac{1}{k}\sum_{i=1}^{k} E\,(X_i-Y_i)^2 = \frac{1}{k}E\left(\sum_{i=1}^{k}(X_i-Y_i)^2\right) = \frac{1}{k}E\,\|\underline{X}-\underline{Y}\|^2$$

$$= \frac{1}{k}E\,\|\underline{U}-\underline{V}\|^2 = \frac{1}{k}E\left(\sum_{i=1}^{k}(U_i-V_i)^2\right) = \frac{1}{k}\sum_{i=1}^{k} E\,(U_i-V_i)^2$$

$$= \frac{1}{k}\sum_{i=1}^{k} E\,(U_i-Q_i(U_i))^2 = \frac{1}{k}\sum_{i=1}^{k} D_{U_i,Q_i}$$

where D_{U_i,Q_i} is the distortion Q_i applied to U_i.

This is why we restrict to orthogonal transforms.

OPTIMAL DESIGN AND THE OPTA FUNCTION

The parameters of a transform code to be chosen: dimension k, k×k orthogonal matrix T, and scalar quantizers $Q_1,\ldots,Q_k$.

Given a rate R and a dimension k, let us find the least possible distortion D. That is, let us find the OPTA function

$\delta_{tr}(k,R) \triangleq$ least dist'n of transform codes with dimension k and rate R or less,

Let us also discover how to design optimal transform codes, i.e. codes that attain the OPTA function.

The derivation is done in two steps.

Step 1: Optimal quantizers for a given transform.

Given dimension k, rate constraint R, and transform T, choose the scalar quantizers $Q_1,\ldots,Q_k$ to minimize D. This is done in two steps.

(a) Given $R_1,\ldots,R_k$ such that $\frac{1}{k}\Sigma_{i=1}^{k}R_i \le R$, choose $Q_1,\ldots,Q_k$ to have rates $R_1,\ldots,R_k$ or less, respectively, and to minimize distortion.

(b) Choose $R_1,\ldots,R_k$ to minimize distortion assuming $Q_1, \ldots, Q_k$ are chosen as in (a). This is called a *bit or rate allocation*.

Step 2: Optimal transform.

Choose transform T to minimize D assuming $R_1,\ldots,R_k$ and $Q_1,\ldots,Q_k$ are chosen as in 1) for T.

Step 1(a): Given k, T and $R_1,\ldots,R_k$ such that $\frac{1}{k}\Sigma_{i=1}^{k} R_i = R$, choose $Q_1,\ldots,Q_k$ to have rates $R_1,\ldots,R_k$, respectively, and to minimize distortion.

Since

$$D = \frac{1}{k}\sum_{i=1}^{k} D_{U_i,Q_i},$$

for each i we choose Q_i to minimize D_{U_i,Q_i} subject to $R(Q_i) \le R_i$. Then

$$D_{U_i,Q_i} = \delta_{sq,U_i}(R_i) \triangleq \text{OPTA for scalar quantization of } U_i.$$

The OPTA for U_i is either the fixed-rate or the variable-rate OPTA, depending on whether we want to optimize transform coding for fixed or variable rate coding.

It follows that the least distortion attainable for the given k, T, and $R_1,\ldots,R_k$ is

$$D = \frac{1}{k}\sum_{i=1}^{k} \delta_{sq,U_i}(R_i)$$

Step 1(b): Bit/rate allocation

We now seek to find the $R_1,\ldots,R_k$ that minimizes

$$\frac{1}{k}\sum_{i=1}^{k} \delta_{sq,U_i}(R_i)$$

subject to the constraint

$$\frac{1}{k}\sum_{i=1}^{k} R_i \le R.$$

This is called a bit or rate allocation, i.e. an allocation of kR bits among the scalar quantizers for the individual coefficients. If some are given rate larger than R, then others must be given rate less than R.

Recall the original idea of transform coding, namely, that some coefficients will be smaller on the average than others, and that these will be encoded at lower rates than those that are larger on the average. If there is to be a net gain, then it will depend on a successful bit allocation.

For fixed-rate coding, the values of $\delta_{sq,U_i}(R_i)$ can be tabulated for the various values of R_i and a numerical optimization algorithm can be applied to find the best choice of the R_i's. See Gersho and Gray, pp. 226-235. Similar methods may be used for variable-rate coding.

We now consider the high-resolution case. In this case, one may solve analytically for the best bit allocation.

HIGH-RESOLUTION ANALYSIS

Assume R is so large that the optimal R_i's are so large that we may use the approximations

$$\delta_{sq,U_i}(R_i) \cong \frac{1}{12}\,\sigma_i^2 \alpha_i\, 2^{-2R_i},$$

where σ_i^2 = variance of U_i, and
for fixed-rate coding

$$\alpha_i = \beta_{U_i} = \frac{1}{\sigma_i^2}\left(\int f_{U_i}^{1/3}(u)\,du\right)^3$$

and for variable-rate coding

$$\alpha_i = \frac{1}{\sigma_i^2}\,2^{2h(U_i)}, \quad \text{where } h(U_i) = -\int_{-\infty}^{\infty} f_{U_i}(u)\,\log_2 f_{U_i}(u)\,du$$

We now have

$$D \cong \frac{1}{k}\sum_{i=1}^{k} \frac{1}{12}\,\sigma_i^2 \alpha_i\, 2^{-2R_i}$$

And the least dist'n for dimen. k, transform T, and rate R or less is approx'y

$$\min_{R_1\ge 0,\ldots,R_k\ge 0:\ \frac{1}{k}\sum_{i=1}^{k} R_i \le R} \frac{1}{k}\sum_{i=1}^{k} \frac{1}{12}\,\sigma_i^2 \alpha_i\, 2^{-2R_i}$$

A USEFUL OPTIMIZATION RESULT

Lemma: Let $g_1(R), \ldots, g_k(R)$ be functions that are continuous and strictly increasing functions, and let $R_1,\ldots,R_k$ minimize

$$\sum_{i=1}^{k} g_i(R_i) \text{ subject to } \frac{1}{k}\sum_{i=1}^{k} R_i \le R.$$

Then

$$g_i'(R_i) = g_j'(R_j) \text{ for all } i,j.$$

That is, at the optimum choice of $R_1,\ldots,R_k$, the slopes of the g_i functions are the same.

(This is the basis of the Lagrange multiplier method.)