Homework 3 Solutions ***DRAFT*** **EECS 651** **Jan. 31, 2001**

1. *(Problem 11.12, p. 403, from Gersho and Gray). A 2-codeword 4-dimensional VQ is to be designed. The distortion measure is Hamming: $d(x,y) = 1$ if $x \neq y$, $d(x,y) = 0$ if $x=y$. The distortion between vectors is the average of the Hamming distortion between their components. Apply the LGB algorithm to the training sequence below*

   1111, 1110, 1110, 0001, 1001, 0001, 1000, 0010, 0001, 1101

   Start with initial codebook $C_1 = \{\underline{w}_1, \underline{w}_2\} = \{1100, 0011\}$. You will have to modify the LBG algorithm to suit this new distortion measure. That is, assume that in case of a tie in the distortion between an input vector and two codewords, the training vector is assigned to $\underline{w}_1$. Also assume that in the case of a tie in the centroid computation that a 0 is chosen.

   Note that the distortion measure between vectors is

   $$d(\underline{x},\underline{y}) = \frac{1}{4}\, d_H(\underline{x},\underline{y})$$

   where $d_H(\underline{x},\underline{y})$ = # places in which $\underline{x}$ and $\underline{y}$ disagree = Hamming distance between $\underline{x}$ and $\underline{y}$. As mentioned in my email, we must first find optimality criteria.

   Optimality Property 1. Given a codebook $C = \{\underline{w}_1, \underline{w}_2\}$ the best partition $S = \{S_1,S_2\}$ has

   $$S_1 = \{\underline{x}:\ d_H(\underline{x},\underline{w}_1) \le d_H(\underline{x},\underline{w}_2\} \text{ and } S_2 = \{\underline{x}:\ d_H(\underline{x},\underline{w}_2) < d_{HH}(\underline{x},\underline{y}_1\}$$

   In other words a vector $\underline{x}$ is quantized to the codevector that is closest in Hamming distance. Note that we don't need the 1/4's and that we have broken "ties" in favor of $\underline{w}_1$. Also, note that there is no reason for a component of a codevector to have a value other than 0 or 1, because such a value will always cause distortion 1. So from now on we make our codewords consist of 0's and 1's.

   Optimality Property 2: Given a partition $S = \{S_1,S_2\}$, from basic estimation principles, the best codebook is $C = \{\underline{w}_1, \underline{w}_2\}$, where $\underline{w}_i$ is the vector such that $E[d_H(\underline{X},\underline{w}_i)|\underline{X} \in S_i]$ is minimized. Note that $E[d_H(\underline{X},\underline{w}_i)|\underline{X} \in S_i]$ is the expected number of places where $\underline{X}$ and $\underline{w}_i$ differ given that $\underline{X} \in S_i$. The resulting $\underline{w}_i$ is again considered to be a "centroid".

   In a training sequence design method instead of design we replace 2 with
   2'. Given a partition $S = \{S_1,S_2\}$ and a training sequence $\{\underline{t}_1, \ldots,\underline{t}_N\}$, the best codebook is

   $C = \{\underline{w}_1, \underline{w}_2\}$ where $\underline{w}_i$ is the "centroid" vector such that $\frac{1}{N}\sum_{j=1}^{N} d_H(\underline{t}_j,\underline{y}_i)$ is minimized. Specifically, $w_{i,j}$ is 0 if there are at least as many zero's in the jth position of training vectors as ones and $w_{i,j} = 1$, if there are more ones in the jth position of training vectors than zero's. (Note that we broke ties as specified earlier.)

   The following table shows the operation of the training sequence method. Each row (below the training vectors) shows an "old codebook" on the left. Then below each training vector is the index of the old codeword to which it is closest in Hamming distance, with ties broken in favor of $\underline{w}_1$. This is, in effect, the partitioning step. Below that is the Hamming distance of this closest codeword. Next the Hamming distortion resulting from this codebook and partition is shown (the 1/4 factors have been omitted). Finally, a new codebook computed as in 2' is given. The new codebook becomes the old codebook on the next line and the process is repeated until no improvement results.

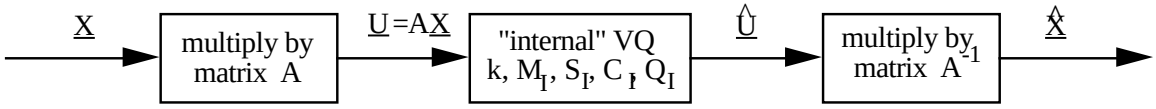
old codebook		training vectors										dist'n old cdbk new partn	new codebook	
$\underline{w}_1$	$\underline{w}_2$	1111	1110	1110	0001	1001	0001	1000	0010	0001	1101		$\underline{c}_1$	$\underline{c}_2$
1100	0011	1	1	1	2	1	2	1	2	2	1		1100	0001
		2	1	1	1	2	1	1	1	1	1	12/10		
1100	0001	1	1	1	2	2	2	1	2	2	1		1110	0001
		2	1	0	0	1	0	1	2	0	1	8/10		
1110	0001	1	1	1	2	2	2	1	1	2	1		1110	0001
		1	0	0	0	1	0	2	2	0	2	8/10		

(a) The final codebook is: {1110, 0001}, because after it produced this, it continued to produce it as the next step.

(b) The final average Hamming distortion is $D = \frac{1}{4}\frac{8}{10} = \frac{1}{5}$.

(Note that average Hamming distance is $\frac{8}{10} = \frac{4}{5}$.)

2. *Consider the vector quantizer described by the following block diagram. (This is a kind of vector generalization of Problem 7 of the previous homework assignment.)*



The source random vector is $\underline{X} = (X_1,...,X_k)^t$ *with pdf* $p_{\underline{X}}(\underline{x})$. *The matrix* A *is a* $k \times k$ *orthogonal matrix, which means it has the properites that* $A^{-1} = A^t$, *the rows are orthonormal, the columns are orthonormal, and* $||A\underline{x}|| = ||\underline{x}||$ *for any* $\underline{x}$ *(each of these properties implies the others). From the diagram we see that* $\underline{U} = A\underline{X}$, $\hat{\underline{U}} = Q_I(\underline{U})$, *and* $\hat{\underline{X}} = A^{-1}\hat{\underline{U}}$.

(a) *Find the codebook* C, *partition* S, *quantization rule* Q, *and rate of the overall quantizer in terms of the matrix* A *and the corresponding properties of the internal VQ.*

Codebook: $\mathbf{C = A^{-1} C_I = \{A^{-1}\underline{w}_{I,i} : i = 1,...,N_I\}}$

Partition: $\mathbf{S = A^{-1} S_I = \{A^{-1}S_{I,i} : i = 1,...,N_I\}}$

Quantization rule: $\mathbf{Q(\underline{x}) = A^{-1} Q_I(A\underline{x})}$

Rate: $\mathbf{R} = \frac{\log_2 M_I}{k} = \mathbf{R_I}$

(b) *Show that the MSE distortion of the overall quantizer operating on* $\underline{X}$ *equals the distortion of the internal quantizer operating on* $\underline{U}$.

The distortion of the overall quantizer operating on $\underline{X}$ is

$$\mathbf{D_X(C)} = \frac{1}{k} E\,||\underline{X}-Q(\underline{X}))||^2 = \frac{1}{k} E\,||A^{-1}\underline{U}-A^{-1}Q_I(A\underline{X})||^2 = \frac{1}{k} E\,||A^{-1}(\underline{U}-Q_I(U))||^2$$

$$= \frac{1}{k} E\,||\underline{U}-Q_I(U)||^2 = \mathbf{D_U(C_I)}.$$

(c) *Show that if the internal VQ is optimal for $\underline{U}$ (meaning that for its size and dimension it has smallest MSE), then the overall VQ is optimal for $\underline{X}$, regardless of which orthogonal matrix is chosen. (The converse is also true, namely, if the overall VQ is optimal for $\underline{X}$, then the internal is optimal for $\underline{U}$, but you don't have to show it.)*

Proof by contradiction. Suppose C_I is an optimal VQ with rate R for $\underline{U}$. Also suppose C is not optimal for $\underline{X}$. Then there must be a better VQ C' with rate R such that $D_{\underline{X}}(C') < D_{\underline{X}}(C)$. Let $C'_I = AC'$ be a code of rate R for $\underline{U}$. Then

$$D_U(C'_I) = D_{\underline{X}}(C') < D_{\underline{X}}(C) = D_U(C_I),$$

which contradicts the optimality of C_I for $\underline{U}$. Thus it must be that C is optimal for $\underline{X}$.

(d) *In conventional "transform coding", such as JPEG, a great deal of attention is paid to choosing the orthogonal matrix. Why is this? (Property (b) seems to be saying that it doesn't matter.)*

In ordinary transform coding, the internal quantizer consists of a bank of scalar quantizers. In this case it matters greatly what transform is chosen. On the other hand a k-dimensional internal VQ could implicitly include whatever transformation one would like to have, so the transform A has no effect on the best possible performance.

(e) *Show that if $\underline{X}$ is multiplied by a constant $1/a > 0$ before being multiplied by the matrix A and if the output of the inverse matrix multiplier is multiplied by a in producing $\hat{\underline{X}}$, then the distortion of the overall quantizer on $\underline{X}$ is a^2 times the distortion of the internal quantizer on $\underline{U}$.*

Let Q' be the new quantization rule and Q be the rule found in (a). Then

$$Q'(\underline{x}) = a\, Q\left(\tfrac{1}{a}\underline{x}\right) = a\, A^{-1} Q_I\left(\tfrac{1}{a} A\underline{x}\right)$$

$$D_X(Q') = \frac{1}{k} E\, \|\underline{X}-Q'(\underline{X})\|^2 = \frac{1}{k} E\, \|\underline{X}-aQ\left(\tfrac{1}{a}\underline{X}\right)\|^2$$

$$\mathbf{D_X(Q')} = \frac{1}{k} E\, \|\underline{X}-Q'(\underline{X}))\|^2 = \frac{1}{k} E\, \|aA^{-1}\underline{U}-aA^{-1}Q_I(\tfrac{1}{a}A\underline{X})\|^2 = \frac{1}{k} a^2\, E\, \|A^{-1}(\underline{U}-Q_I(U))\|^2$$

$$= \frac{1}{k} a^2\ E\, \|\underline{U}-Q_I(U)\|^2 = \mathbf{a^2\, D_U(Q_I)}.$$

(f) *Assuming that the internal quantizer has point density $\lambda_I(\underline{x})$ and inertial profile $m_I(\underline{x})$, find the point density $\lambda(\underline{x})$ and inertial profile $m(\underline{x})$ of the overall quantizer in terms of A and the internal point density and inertial profile.*

Here we do not assume the factor "a" is used. The point density is

$$\lambda(\mathbf{\underline{x}}) \cong \frac{1}{N\, \mathrm{vol}(S_{\underline{x}})} = \frac{1}{M_I\, |A^{-1}S_{I,A\underline{x}}|} = \frac{1}{M_I\, |S_{I,A\underline{x}}|} = \lambda_{\mathbf{I}}(\mathbf{A\underline{x}})$$

where $S_{\underline{x}}$ denotes the cell of the overall quantizer containing $\underline{x}$ and $S_{I,\underline{u}}$ denotes the cell of the internal quantizer containing $\underline{u}$.

The inertial profile is

$$\mathbf{m(\underline{x})} = M(S_{\underline{x}}) = M(A^{-1}S_{I,A\underline{x}}) = M(S_{I,A\underline{x}}) = \mathbf{m_I(A\underline{x})}$$

3. *Consider a wide-sense stationary, first-order autoregressive source of the form*

$$X_n = \rho X_{n-1} + Z_n$$

where the Z_n's are IID with zero means and where Z_n is uncorrelated with $X_{n-1}, X_{n-2}, \ldots$. Show that

$$E\,Z^2 = E\,X^2\,(1-\rho^2)\,.$$

We have

$$E\,X_n^2 = E\,(\rho\,X_{n-1} + Z_n)^2 = \rho^2\,E\,X_{n-1}^2 + 2\rho\,E\,X_{n-1}\,Z_n + E\,Z_n^2$$

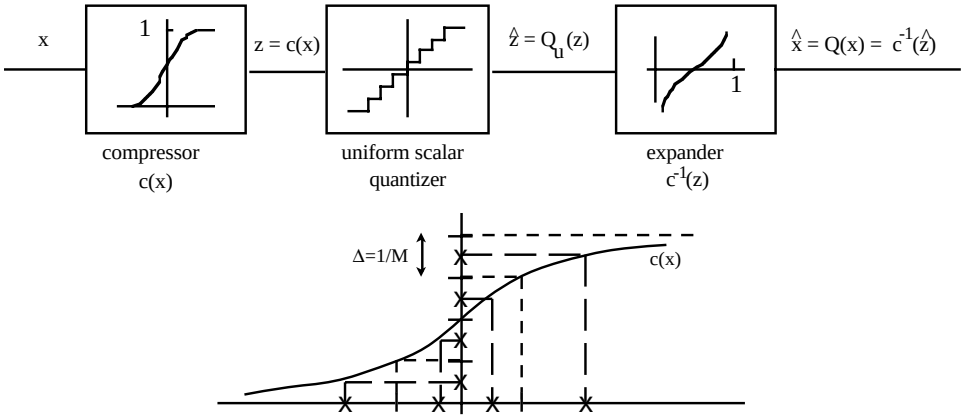
because the process is stationary $E\,X_n^2 = E\,X_{n-1}^2 = E\,X^2$ and $E\,Z_n^2 = E\,Z^2$. Since X_{n-1} and Z_n are uncorrelated and $E\,Z_n = 0$, we have $E\,X_{n-1}\,Z_n = 0$. Therefore, the above becomes

$$E\,X^2 = \rho^2\,E\,X^2 + E\,Z^2$$

which implies

$$E\,Z^2 = E\,X^2\,(1-\rho^2)$$

4. *Consider the scalar quantizer shown below, called a compander, that quantizes by preceding the encoder of an M level uniform scalar quantizer with support [0,1] with a memoryless nonlinear function c(x). At the decoder, the output of the decoder for the uniform scalar quantizer is followed by the inverse of c. The levels and thresholds of the uniform scalar quantizer are distributed evenly over the interval [0,1]. The function c is nonnegative and montonically increasing, and it maps (-∞,∞) into [0,1]. The plot below the block diagram may help you to visualize the operation of the compander.*



(a) *Find formulas for the levels $w_1,\ldots,w_M$ and thresholds $t_0,\ldots,t_M$ of the compander in terms of the function c.*

The ith level of the uniform scalar quantizer is $v_i = \frac{i}{M} - \frac{1}{2M}$ and the ith level of the compander is

$$w_i = c^{-1}\left(\frac{i}{M} - \frac{1}{2M}\right),\quad i = 1,\ldots,M$$

The ith threshold of the uniform scalar quantizer is $\frac{i}{M}$, $u_i = i = 1,\ldots,M-1$, $u_0 = 0$, $u_M = 1$. The ith threshold of the compander is

$$t_i = c^{-1}\left(\frac{i}{M}\right),\quad i = 1,\ldots,M \text{ and } t_0 = -\infty,\ t_M = \infty.$$

I'm using the convention that the thresholds are defined so that $Q(x) = w_i$ if $t_{i-1} < x < t_i$.

(b) *Assuming M is large, find an approximate expression for the distortion of this quantizer in terms of M, the function c, and the probability density of X. Simplify as much as possible (Hint: It should be an integral expression.)*

Let's use Bennett's integral. The dimension $k = 1$. The point density is

$$\lambda(x) = \frac{1}{M}\frac{1}{t_i - t_{i-1}} \quad \text{if } t_{i-1} < x < t_i$$

Since M is large, we can ssume $\frac{1}{M}$ and $(t_i - t_{i-1})$ are small and we can use the approximation

$$c'(x) \cong \frac{1/M}{t_i - t_{i-1}}$$

From this it follows that

$$\lambda(x) \cong c'(x)$$

To find the inertial profile we note that when M is large, the cells will be small and $c(x)$ will be approximately linear across one cell. From this it follows that w_i will be approximately in the center of its cell. From this it follows that the inertial profile is

$$m(x) \cong \frac{1}{12}$$

Substituting the point density and inertial profile into Bennett's integral gives

$$\mathbf{D} \cong \frac{\mathbf{1}}{\mathbf{12}}\frac{\mathbf{1}}{\mathbf{M^2}}\int_{-\infty}^{\infty}\frac{\mathbf{1}}{\mathbf{(c'(x))^2}}\,\mathbf{f(x)\,dx}$$

(c) *Show that any scalar quantizer can be implemented with a compander, provided its levels lie within its cells.*

Let $w_1,...,w_M$ and $t_0,...,t_M$ be the thresholds of an arbitrary scalar quantizer with $t_{i-1} < w_i < t_i$, $i = 1,...,M$. We will design a compander that implements this quantizer.

Let $c(x)$ be any continuous, monotonically increasing function that goes to zero as $x \to -\infty$, goes to one as $x \to \infty$, and such that $c(x)$ passes through the following points:

$$(w_1,\frac{1}{2M}),\ (t_1,\frac{1}{M}),\ (w_2,\frac{3}{2M}),\ (t_2,\frac{2}{M}),\ (w_3,\frac{5}{2M}),\ ...,\ (t_{M-1},\frac{M-1}{M}),\ (w_M,1-\frac{1}{2M})\ ,$$

i.e. c is chosen so that

$$c(w_i) = \frac{i}{M} - \frac{1}{2M} \quad \text{and} \quad c(t_t) = \frac{i}{M}$$

Note that each point is to the right and above the previous points. Therefore, there does indeed exist a continuous function that passes through these points. We can also choose it so that it goes to zero as $x \to -\infty$ and so that it goes to one as $x \to \infty$. Such a compander will implement the given quantizer because if $t_{i-1} < x < t_i$ then compressor mapping produces $c(x)$ with $\frac{i-1}{M} < c(x) < \frac{i}{M}$, so that the uniform quantizer choose level $\frac{i}{M} - \frac{1}{2M}$ and the expander mapping produces $w_i = c^{-1}(\frac{i}{M} - \frac{1}{2M})$. The latter happens because c was chosen so that $c(w_i) = \frac{i}{M} - \frac{1}{2M}$.

5. *(a) Use Bennett's integral and the results of Parts b and e of Problem 2 to predict the MSE of JPEG applied to the image 'lena' with quality factor 1. To do this you will need to know that JPEG has the form shown in Problem 2 with the internal quantizer consisting of 64 uniform scalar quantizers with step sizes shown in the table that was distributed and posted on the website. The orthonormal transform is preceded by multiplying by 1/a = 16 and the inverse transform is postmultiplied by a. (You might want to use Matlab, Excel, or write a computer program to avoid a lot of repititious calculations.)*

JPEG is a coder of the kind described in Part e of Problem 2. Therefore,

$$D_X(Q) = a^2 D_U(Q_I),$$

where Q denotes the overall effect of JPEG quantization, $a = 1/16$, $\underline{U} = \mathrm{DCT}(\underline{X})$, and Q_I denotes the quantization of the DCT coefficients in $\underline{U} = (U_1,...,U_{64})$. We now compute

$$D_U(Q_I) = \frac{1}{64} E\|\underline{U}-\hat{\underline{U}}\|^2 = \frac{1}{64} E \sum_{j=1}^{64} (U_j - Q_j(U_j))^2 = \frac{1}{64} \sum_{j=1}^{64} E\,(U_j - Q_j(U_j))^2$$

where Q_j denotes the quantization done by JPEG to the jth coefficient. According to how JPEG operates Q_j is a uniform scalar quantizer with step size $\Delta_j = M_j$ where M_j is the jth element of the quantization matrix. As derived in class using Bennett's integral

$$E\,(U_j - Q_j(U_j))^2 \cong \frac{\Delta_j^2}{12} = \frac{M_j^2}{12}$$

Therefore, our prediction is

$$\mathbf{D_X(Q)} = a^2 D_U(Q_I) \cong \frac{1}{256} \frac{1}{64} \sum_{j=1}^{64} \frac{M_j^2}{12} = \mathbf{1.464}$$

(b) *Compare to the actual distortion of JPEG running on 'lena' with quality factor 1.*

The actual MSE on lena is **D = 0.0925**.

The considerable difference between this and the actual value is due to the fact that most of the M_j's are not small relative to the standard deviation of the variable being quantized. Indeed, most of the coefficients have quantization step sizes much much larger than the typical value of the coefficient being quantized. The formula for D predicts the distortion for coefficient U_j to be $M_j^2/12$ when in fact because $U_j << M_j$, most values of the time U_j is quantized to 0 and the distortion is approximately $E[(U_j)^2] << M_j^2/12$.