CHAPTER 3: TRANSFORMS

1. Why use Transform Coding?
2. How much can we gain from TC?
3. What would be a practical implementation of TC?

Transform Coding

Transform Coding is nothing but apply a transformation to the input symbols before quantization. The reason behind this transformation is to eliminate redundancy. At the end we want to lower the transmitted bit rate.

Let's formulate the mathematical problem of this new model and then we will see how we can improve the transmission by lowering the bit rate.

Suppose the input or our system is composed of all the symbols $x = \{x_1, x_2, ..., x_N\}$
A transformation will output $y = \{y_1, y_2, ..., y_N\}$.
Let's take the input and output as if they were vectors. Then:

$\mathbf{y} = \mathbf{K}\mathbf{x}$ where $\mathbf{x} = [x_1, x_2, ..., x_N]^T$, $\mathbf{y} = [y_1, y_2, ..., y_N]^T$, and $\mathbf{K} = \begin{bmatrix} k_{11} & .. & k_{1N} \\ .. & & .. \\ k_{N1} & .. & k_{NN} \end{bmatrix}$

We know that for a source with arbitrary PDF we have:
$D(R_i) = \lambda \sigma_{x_i}^2 2^{-2R_i}$ where $\sigma_{x_i}^2$ is the variance of the input.

The variance is defined as $\sigma_{x_i}^2 = E\left[(x_i - \mu_x)^2\right]$

Let's obtain a relation between the variances of the input and the output.

$$Cov(X, X) = E\left[(\mathbf{x} - \mu_x)(\mathbf{x} - \mu_x)^T\right]$$

$$= E\left[\begin{bmatrix} x_1 - \mu_x \\ \\ x_N - \mu_x \end{bmatrix}\begin{bmatrix} x_1 - \mu_x & & x_N - \mu_x \end{bmatrix}\right]$$

$$= E\begin{bmatrix} (x_1 - \mu_x)(x_1 - \mu_x) & & (x_1 - \mu_x)(x_N - \mu_x) \\ \\ (x_N - \mu_x)(x_1 - \mu_x) & & (x_N - \mu_x)(x_N - \mu_x) \end{bmatrix}$$

$$= \begin{bmatrix} E[(x_1 - \mu_x)(x_1 - \mu_x)] & & E[(x_1 - \mu_x)(x_N - \mu_x)] \\ \\ E[(x_N - \mu_x)(x_1 - \mu_x)] & & E[(x_N - \mu_x)(x_N - \mu_x)] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{x_1}^2 & E[(x_1 - \mu_x)(x_N - \mu_x)] \\ E[(x_N - \mu_x)(x_1 - \mu_x)] & \sigma_{x_N}^2 \end{bmatrix}$$

We can see that this matrix is symmetric.

Also:
$Cov(Y,Y) = E\left[(\mathbf{y} - \mu_y)(\mathbf{y} - \mu_y)^T\right]$. However we know that $\mathbf{y} = \mathbf{Kx}$. Then:
$Cov(Y,Y) = E\left[(\mathbf{Kx} - \mathbf{K}\mu_x)(\mathbf{Kx} - \mathbf{K}\mu_x)^T\right]$ By properties of matrices we have
$Cov(Y,Y) = E\left[\mathbf{K}(\mathbf{x} - \mu_x)(\mathbf{K}[\mathbf{x} - \mu_x])^T\right] = E\left[\mathbf{K}(\mathbf{x} - \mu_x)([\mathbf{x} - \mu_x]^T \mathbf{K}^T)\right]$. Finally
We found the relationship between covariance matrix of the input and output:

$Cov(Y,Y) = \mathbf{K}Cov(X,X)\mathbf{K}^T$

We will find out later, that in order to ideally reduce the transmission bit rate, the $Cov(Y,Y)$ matrix should be diagonal.

Therefore, ideally we need a Transform Coding that converts any input with $Cov(X,X)$ into an output with diagonal $Cov(Y,Y)$.

This is a common mathematical problem called SVD. And it states that if we have a symmetrical matrix **A**, we can find a diagonal matrix Λ that holds the relationship $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^T$ **U** is a orthonormal matrix (where $\mathbf{UU^T=I}$)

In our particular case we know that the covariance matrix of the input is symmetric.
Therefore:
$C_{xx} = \mathbf{U}\Lambda\mathbf{U}^T$
If we want the covariance matrix of the output to be symmetric, we can have:
$C_{xx} = \mathbf{U}C_{yy}\mathbf{U}^T$

The relationship of the covariance matrices are:
$C_{yy} = \mathbf{K}C_{xx}\mathbf{K}^T$

Operating we have:
$\mathbf{U}^T C_{xx} = \mathbf{U}^T\mathbf{U}C_{yy}\mathbf{U}^T = C_{yy}\mathbf{U}^T$
$\mathbf{U}^T C_{xx}\mathbf{U} = C_{yy}\mathbf{U}^T\mathbf{U} = C_{yy}$ or
$C_{yy} = \mathbf{U}^T C_{xx}\mathbf{U}$

Therefore the Transform Coding that obtains a diagonal covariance matrix at the output should hold this equation:

$\mathbf{K} = \mathbf{U}^T$ where $\mathbf{U}$ is the solution of the SDV theorem.

This process is the "Karhunnen-Loeve Transform" or also called "Principal Component Analysis".

Performance Gain Analysis

For a source with arbitrary PDF, after quantization we have:

$$\boxed{D(R_i) = \lambda \sigma_{x_i}^2 2^{-2R_i}}$$

The overall distortion for each one of the inputs is:

$$D_{overall} = \sum_{i=1}^{N} D(R_i) = \sum_{i=1}^{N} \lambda \sigma_{x_i}^2 2^{-2R_i} = \lambda \sum_{i=1}^{N} \sigma_{x_i}^2 2^{-2R_i}$$

(In the book we have $D_{TC} = \frac{1}{N}\sum_{k=1}^{N} D(R_k) = \frac{1}{N}\sum_{k=1}^{N} \varepsilon_{t,k}^2 \sigma_{t,k}^2 2^{-2R_k}$)

Optimal Bit Allocation

We want to make this minimum with the constrain of:

$\sum_{i=1}^{N} R_i \le R_T$ where R_T is the total bit rate (in the book this is NR where R is the average bit rate)

Using the Lagrange multiplier:

$$J = \lambda \sum_{i=1}^{N} \sigma_{x_i}^2 2^{-2R_i} + \eta\left(\sum_{i=1}^{N} R_i - R_T\right)$$

Derivation with respect R_i

$$\frac{\partial J}{\partial R_i} = \frac{\partial\left[\lambda \sum_{i=1}^{N} \sigma_{x_i}^2 2^{-2R_i} + \eta\left(\sum_{i=1}^{N} R_i - R_T\right)\right]}{\partial R_i} = 0$$

$$\frac{\partial J}{\partial R_i} = -2\ln 2 \lambda \sigma_{x_i}^2 2^{-2R_i} + \eta = 0$$

Therefore we can have $\eta(R_i)$ as

$\eta(R_i) = 2\ln 2 \lambda \sigma_{x_i}^2 2^{-2R_i}$ or because $D(R_i) = \lambda \sigma_{x_i}^2 2^{-2R_i}$

$\boxed{\eta(R_i) = 2\ln 2 D(R_i)}$ or:

$\boxed{D(R_i) = \frac{\eta}{2\ln 2}}$ This means the distortion for each input is the same.

Let's find the bit rate per each symbol function of the total bit rate.
We found:
$\eta(R_i) = 2\ln 2\lambda\sigma_{x_i}^2 2^{-2R_i}$ if we multiply for all *i's* we have:

$\eta^N(R_i) = \left(2\ln 2\lambda\right)^N \prod_{i=1}^{N} \sigma_{x_i}^2 2^{-2\sum_{i=1}^{N} R_i}$ being $\sum_{i=1}^{N} R_i = R_T$ Therefore:

$\eta^N(R_T) = \left(2\ln 2\lambda\right)^N \prod_{i=1}^{N} \sigma_{x_i}^2 2^{-2R_T}$ taking Nth root we have

$$\eta(R_T) = \left(2\ln 2\right)\lambda\left(\prod_{i=1}^{N} \sigma_{x_i}^2\right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}$$

From this expression we can find some other interesting relations.
Let's find $R_i(R_T)$

$\eta(R_i) = 2\ln 2\lambda\sigma_{x_i}^2 2^{-2R_i}$ and $\eta(R_T) = \left(2\ln 2\right)\lambda\left(\prod_{i=1}^{N} \sigma_{x_i}^2\right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}$ Therefore

$$2\ln 2\lambda\sigma_{x_i}^2 2^{-2R_i} = \left(2\ln 2\right)\lambda\left(\prod_{i=1}^{N} \sigma_{x_i}^2\right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}$$

$$\sigma_{x_i}^2 2^{-2R_i} = \left(\prod_{i=1}^{N} \sigma_{x_i}^2\right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}$$

$$\frac{\sigma_{x_i}^2}{\left(\prod_{i=1}^{N} \sigma_{x_i}^2\right)^{\frac{1}{N}}} = 2^{-2\frac{R_T}{N}} 2^{+2R_i}$$

$$\log_2 \frac{\sigma_{x_i}^2}{\left(\prod_{i=1}^{N} \sigma_{x_i}^2\right)^{\frac{1}{N}}} = -2\frac{R_T}{N} + 2R_i \quad \text{Finally}$$

$$\boxed{R_i(R_T) = \frac{1}{2}\log_2 \frac{\sigma_{x_i}^2}{\left(\prod_{i=1}^{N} \sigma_{x_i}^2\right)^{\frac{1}{N}}} + \frac{R_T}{N}}$$

Also we can find the distortion in function of R_T

$$D(R_i) = \frac{\eta}{2\ln 2} = \frac{\left(2\ln 2\right)\lambda\left(\prod_{i=1}^{N} \sigma_{x_i}^2\right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}}{2\ln 2}$$

$$\boxed{D(R_T) = \lambda \left(\prod_{i=1}^{N} \sigma_{x_i}^2 \right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}}$$

Summarizing we have:

$$\boxed{D(R_i) = \lambda \sigma_{x_i}^2 \, 2^{-2R_i} = \frac{\eta}{2\ln 2}}$$

$$\boxed{R_i(R_T) = \frac{1}{2}\log_2 \frac{\sigma_{x_i}^2}{\left(\prod_{i=1}^{N} \sigma_{x_i}^2 \right)^{\frac{1}{N}}} + \frac{R_T}{N}}$$

$$\boxed{D(R_T) = \lambda \left(\prod_{i=1}^{N} \sigma_{x_i}^2 \right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}}$$

Introducing the Transform Coding with Bit Allocation

Let’s see what happens if having bit allocation we compare the distortion rate when having Transform Coding and if we do not have it.

Without Transform coding:

$$D_{NTC}(R_T) = \lambda \left(\prod_{i=1}^{N} \sigma_{x_i}^2 \right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}$$

With Transform Coding we have:

$$D_{TC}(R_T) = \lambda \left(\prod_{i=1}^{N} \sigma_{y_i}^2 \right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}} \text{ where } y_i = TC(x_i)$$

<u>If the input is stationary</u> (each symbol/input has the same variance) we have that:

$$D_{NTC}(R_T) = \lambda \left(\prod_{i=1}^{N} \sigma_{x_i}^2 \right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}} = \lambda \sigma_{x_i}^2 \, 2^{-2\frac{R_T}{N}} \text{ and we can do a mathematical trick}$$

$$D_{NTC}(R_T) = \lambda \sigma_{x_i}^2 \, 2^{-2\frac{R_T}{N}} = \lambda \frac{1}{N} \sum_{i=1}^{N} \sigma_{x_i}^2 \, 2^{-2\frac{R_T}{N}} \text{ because } \frac{1}{N} \sum_{i=1}^{N} \sigma_{x_i}^2 = \sigma_{x_i}^2$$

Now we will compare both distortions:

$$D_{NTC}(R_T) = \lambda \frac{1}{N} \sum_{i=1}^{N} \sigma_{x_i}^2 \, 2^{-2\frac{R_T}{N}} \text{ and } D_{TC}(R_T) = \lambda \left(\prod_{i=1}^{N} \sigma_{y_i}^2 \right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}$$

$$\frac{D_{NTC}(R_T)}{D_{TC}(R_T)} = \frac{\lambda \frac{1}{N}\sum_{i=1}^{N}\sigma_{x_i}^2 2^{-2\frac{R_T}{N}}}{\lambda\left(\prod_{i=1}^{N}\sigma_{y_i}^2\right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}} = \frac{\frac{1}{N}\sum_{i=1}^{N}\sigma_{x_i}^2}{\left(\prod_{i=1}^{N}\sigma_{y_i}^2\right)^{\frac{1}{N}}}$$

<u>If the transform is orthonormal</u>, then: $\sum_{i=1}^{N}\sigma_{x_i}^2 = \sum_{i=1}^{N}\sigma_{y_i}^2$. Which is an equivalent of the Parseval's theorem for the Fourier Transform.

Therefore:

$$\frac{D_{NTC}(R_T)}{D_{TC}(R_T)} = \frac{\frac{1}{N}\sum_{i=1}^{N}\sigma_{y_i}^2}{\left(\prod_{i=1}^{N}\sigma_{y_i}^2\right)^{\frac{1}{N}}}$$ (Stationary input and orthonormal transform)

Using Jensen's inequality it can be proved that the numerator is always greater than the denominator. The numerator is nothing else but the Arithmetic Mean. The denominator is the Geometric Mean.

Therefore we can summarize:
If we apply an orthogonal Transform to a stationary input and we quantize it under the constrain of bit allocation, we lower the Distortion Rate.

Finding the best Transform Coding

So far we have found that we lower the Distortion Rate if we apply an orthogonal Transform to a stationary input and then quantize it under the constrain of bit allocation. However, what it is the best Transform that achieves the lowest Distortion rate?

We are going to prove that the KLT is the optimal transformation.

Our goal is to lower the expression:

$$D_{NTC}(R_T) = \lambda\left(\prod_{i=1}^{N}\sigma_{y_i}^2\right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}$$

Therefore we have to minimize the product of the variances $\prod_{i=1}^{N}\sigma_{y_i}^2$

Let's imagine that we apply any transformation to the input. In that case the covariance matrix of the output does not have to be diagonal. The only case the covariance matrix of the output is diagonal happens if we applied the KLT to the input.

We know that for any Transform the covariance matrix is symmetric. Therefore using matrix properties we can affirm that for any symmetric matrices the determinant is less or equal than the product of the main diagonal: $|C_{yy}| \leq \prod \sigma_{y_{ii}}^2$ (the equality happens if the Matrix is diagonal).

For any orthonormal transformation the next equation holds: $|C_{y_1y_1}| = |C_{y_2y_2}|$ being y_1, y_2 outputs of two orthonormal Transformations.

Proof:
$C_{y_1y_1} = \mathbf{K}_1 C_{xx} \mathbf{K}^T{}_1$ and $C_{y_2y_2} = \mathbf{K}_2 C_{xx} \mathbf{K}^T{}_2$
Taking determinants:
$|C_{y_1y_1}| = |\mathbf{K}_1||C_{xx}||\mathbf{K}^T{}_1|$
$|C_{y_2y_2}| = |\mathbf{K}_2||C_{xx}||\mathbf{K}^T{}_2|$
Because $\mathbf{K_1}$, $\mathbf{K_2}$ are orthonormal, its determinant is one. Therefore
$|C_{y_1y_1}| = |C_{xx}| = |C_{y_2y_2}|$ Hence it is proved.

Therefore if we apply two transforms to the same input, one being KLT and the other not, we can assure this equations holds:

$$|C_{y_1y_1KLT}| = |C_{y_1y_1NOTKLT}|$$

Now, for a KLT we know that the matrix is diagonal, therefore

$|C_{y_1y_1KLT}| = \prod \sigma_{y_{1ii}}^2$ Because is a diagonal matrix.

Also this is true:

$|C_{y_2y_2NOTKLT}| \leq \prod \sigma_{y_{2ii}}^2$ Because in general is not a diagonal matrix.

Finally we have that:

$$\prod \sigma_{y_{1ii}}^2 = |C_{y_1y_1KLT}| = |C_{y_2y_2NOTKLT}| \leq \prod \sigma_{y_{2ii}}^2$$

Therefore the product of the variances of a KLT transform is lesser or equal the product of the variances of any other transform.

Because Distortion Rate expression is

$D_{TC}(R_T) = \lambda\left(\prod_{i=1}^{N}\sigma_{y_i}^2\right)^{\frac{1}{N}} 2^{-2\frac{R_T}{N}}$ We can assure than:

$$\boxed{D_{KL-TC}(R_T) \leq D_{NOTKL-TC}(R_T)}$$

Therefore the KLT is optimal to minimize the Distortion Rate.