Bit allocation in scalar quantization

- Problem formulation: A block of r.v.'s $X_1, X_2, \ldots, X_k$ is scalar quantized. The $X_i$ might have different distributions.

- We want to minimize the overall MSE

$$E \left[ \sum_{i=1}^{k} (X_i - Q_i(X_i))^2 \right]$$

where $Q_i$ has $N_i$ levels, under the condition that overall no more than $B$ bits are used, i.e.

$$\sum_{i=1}^{k} \log N_i \leq B$$

- Define

$$W_i(b) = \min_{Q_i \cdot r(Q) \leq b} E[|X_i - Q_i(X_i)|^2]$$

(MSE of the optimal $b$-bit quantizer for $X_i$).

- If $b_i$ bits are used to quantize $X_i$ optimally, the overall distortion is

$$D(b) = \sum_{i=1}^{k} W_i(b_i), \quad b = (b_1, \ldots, b_k)$$

Bit allocation problem:

Given the constraint $\sum_{i=1}^{k} b_i \leq B$, find $b = (b_1, \ldots, b_k)$ minimizing $D(b)$.

Note: The $b_i$ are nonnegative integers. If the functions $W_i(b)$ are known, the bit allocation problem can (in principle) be solved by exhaustive search.

- Simplifying assumption: Each $X_i$ has pdf $f_i$ and

$$W_i(b) = \frac{1}{12} \| f_i \|_{1/3} 2^{-2b}, \quad i = 1, \ldots, k$$

Thus we assume that the high-rate approximation to optimal performance holds with equality.

- Let $\sigma_i^2 = \text{Var}(X_i)$ and $\tilde{X}_i = \frac{X_i}{\sigma_i}$. Then $\tilde{X}_i$ has unit variance and its pdf $\tilde{f}_i(x) = \sigma_i f(\sigma_i x)$ satisfies

$$\| f_i \|_{1/3} = \| \tilde{f}_i \|_{1/3} \sigma_i^2$$

Thus

$$W_i(b) = \frac{1}{12} \| \tilde{f}_i \|_{1/3} \sigma_i^2 2^{-2b}$$

$$= h_i \sigma_i^2 2^{-2b}$$

Note: $h_i$ is invariant to scaling and depends only on the shape of the density. E.g., $h_i$ is the same for all normal r.v.'s.
We can solve the bit allocation problem analytically if we relax the condition that the $b_i$ are nonnegative integers.

**Theorem 1 (optimal bit allocation)**

\[
D(b) = \sum_{i=1}^{k} h_i \sigma_i^2 2^{-2b_i}
\]

is minimized subject to \(\sum_{i=1}^{k} b_i \leq B\) if and only if

\[
b_i = \bar{b} + \frac{1}{2} \log_2 \frac{\sigma_i^2}{\rho^2} + \frac{1}{2} \log_2 \frac{h_i}{H}, \quad i = 1, \ldots, k
\]

where

\[
\bar{b} = \frac{B}{k}, \quad \rho^2 = \left(\prod_{i=1}^{k} \sigma_i^2\right)^{1/k}, \quad H = \left(\prod_{i=1}^{k} h_i\right)^{1/k}
\]

**Observation:** The optimal $b$ must satisfy \(\sum_{i=1}^{k} b_i = B\).

**Proof:** Assume \(\sum_{i=1}^{k} b_i < B\). Let

\[
b = B - \sum_{i=1}^{k} b_i
\]

and define \(\hat{b} = (b_1 + b_2, \ldots, b_k)\). Then \(\sum_{i=1}^{k} \hat{b}_i = B\) and

\[
D(b) - D(\hat{b}) = \sum_{i=1}^{k} h_i \sigma_i^2 2^{-2b_i} - \sum_{i=1}^{k} h_i \sigma_i^2 2^{-2\hat{b}_i}
\]

\[
= h_1 \sigma_1^2 (2^{-2b_1} - 2^{-2(b_1+b)}) > 0
\]

since $b > 0$. Thus $b$ cannot solve the optimal bit allocation problem. \(\square\)

**Detour: The Lagrange multiplier method**

Assume $f : \mathbb{R}^k \to \mathbb{R}$ and $g : \mathbb{R}^k \to \mathbb{R}$ are continuously differentiable functions. We want to solve the following constrained minimization problem (CMP):

\[
\text{minimize } f(x)
\]

subject to \(g(x) = 0\)

Let $\nabla u = \left(\frac{\partial}{\partial x_1} u, \ldots, \frac{\partial}{\partial x_k} u\right)$ denote the gradient of a $u : \mathbb{R}^k \to \mathbb{R}$.

**Theorem 2 (Lagrange multiplier)**

If $x^*$ is a solution of the CMP and $\nabla g(x^*) \neq 0$, then there exists $\lambda \in \mathbb{R}$, called the Lagrange multiplier, such that

\[
\nabla f(x^*) + \lambda \nabla g(x^*) = 0
\]

**Remarks:**

- The Lagrange multiplier theorem yields the equations

\[
\frac{\partial}{\partial x_j} \left(f(x) + \lambda g(x)\right)\bigg|_{x=x^*} = 0, \quad j = 1, \ldots, k
\]

\[
g(x^*) = 0
\]

We have $k + 1$ equations for $k + 1$ unknowns $(\lambda, x^*)$.

- Assume the CMP has a solution and $\nabla g(x) \neq 0$ for all $x$ such that $g(x) = 0$. Then the theorem implies that if $(\lambda, x^*)$ uniquely solve these $k + 1$ equations, then $x^*$ is a solution of the CMP.

- In general, the Lagrange multiplier theorem only gives a necessary condition for optimality.
Proof sketch of Lagrange multiplier theorem: Assume \( x^* \) is a solution of the CMP, i.e.,

\[
g(x^*) = 0 \quad \text{and} \quad f(x^*) \leq f(x) \quad \text{for all } x \text{ such that } g(x) = 0
\]

We have to show that if \( \nabla g(x^*) \neq 0 \), then for some \( \lambda \)

\[
\nabla f(x^*) + \lambda \nabla g(x^*) = 0
\]

Let \( S = \{ x : g(x) = 0 \} \). Then \( x^* \in S \) and \( S \) is an \((k-1)\)-dimensional smooth surface in \( \mathbb{R}^k \).

The tangent plane \( T \) of \( S \) at \( x^* \) is the collection of all derivatives \( x'(0) \) of smooth curves \( x(t), t \in \mathbb{R} \) such that \( x(t) \in S \) and \( x(0) = x^* \).

We have shown that

\[
\nabla g(x^*) y = 0 \quad \Rightarrow \quad \nabla f(x^*) y = 0
\]

This means that \( \nabla f(x^*) \) is orthogonal to the subspace

\[
T = \{ y : \nabla g(x^*) y = 0 \}
\]

Note that \( T \) is \((k-1)\)-dimensional because \( \nabla g(x^*) \neq 0 \). Hence its orthogonal complement

\[
T^\perp = \{ x : x^t y = 0 \text{ for all } y \in T \}
\]

is 1-dimensional. Since \( T^\perp \) includes both \( \nabla g(x^*) \) and \( \nabla f(x^*) \), we must have

\[
\nabla f(x^*) = -\lambda \nabla g(x^*)
\]

for some \( \lambda \in \mathbb{R} \). \( \square \)

Lemma 3

If \( \nabla g(x^*) \neq 0 \), then the tangent plane of \( S \) at \( x^* \) is given by

\[
T = \{ y : \nabla g(x^*) y = 0 \}
\]

Now let \( y \in T \). By the lemma, there exists a smooth curve \( x(t) \in S \) such that \( x(0) = x^* \) and \( x'(0) = y \).

Then \( u(t) = f(x(t)) \) is differentiable and since \( x(t) \in S \) and \( x^* \) solves the CMP

\[
u(0) = f(x(0)) = f(x^*) \leq f(x(t)) = u(t)
\]

Thus \( u(t) \) has a minimum at \( t = 0 \), so

\[
0 = u'(t)|_{t=0} = \frac{d}{dt} f(x(t))|_{t=0} = \nabla f(x^*) x'(0) = \nabla f(x^*) y
\]

Illustration of proof for \( k = 2 \)

- Let \( x(t) \) be a smooth curve in \( S \) such that \( x(0) = x^* \).
- We have \( x'(0) \in T \).
- Assume \( \nabla f(x^*) \) is not perpendicular to \( T \).
- In this example, \( \nabla f(x^*) x'(0) < 0 \).
- Using a first order approximation, for small \( t \)

\[
\begin{align*}
f(x(t)) &= f(x(0)) + \frac{d}{dt} f(x(t))|_{t=0} t + o(t) \\
&= f(x^*) + \nabla f(x^*) x'(0) t + o(t)
\end{align*}
\]

where \( o(t)/t \to 0 \) as \( t \to 0 \). Since \( \nabla f(x^*) x'(0) < 0 \), we obtain \( f(x(t)) < f(x^*) \) for \( t > 0 \) small enough.
- Since \( x(t) \in S \), this would contradict the optimality of \( x^* \). Thus \( \nabla f(x^*) x'(0) = 0 \) must hold, i.e., \( \nabla f(x^*) = -\lambda \nabla g(x^*) \) for some \( \lambda \).
Proof of bit allocation formula: We apply the Lagrange multipliers method with
\[
\begin{align*}
  f(b) &= D(b) = \sum_{i=1}^{k} h_i \sigma_i^2 2^{-2b_i}, \\
  g(b) &= \sum_{i=1}^{k} b_i - B
\end{align*}
\]
Define
\[
J(b, \lambda) = \sum_{i=1}^{k} h_i \sigma_i^2 2^{-2b_i} + \lambda \left( \sum_{i=1}^{k} b_i - B \right)
\]
Then for the optimum \( b \) there is a \( \lambda \) such that
\[
\frac{\partial}{\partial b_j} J(b, \lambda) = 0, \quad j = 1, \ldots, k \quad \quad (\ast)
\]
Fix \( \lambda \) and solve (\ast) for \( b \). Then determine \( \lambda \) and \( b = b(\lambda) \) so that the solution satisfies the constraint \( \sum_{i=1}^{k} b_i = B \). If the solution \((\lambda, b)\) is unique, the resulting \( b \) is optimal.

Hence
\[
\hat{\lambda} = B \frac{k}{k} - \frac{1}{2} \log_2 \left( \prod_{i=1}^{k} h_i \right)^{1/k} - \frac{1}{2} \log_2 \left( \prod_{i=1}^{k} \sigma_i^2 \right)^{1/k}
\]
Substituting into
\[
b_j = \frac{1}{2} \log_2 h_j + \frac{1}{2} \log_2 \sigma_j^2 + \hat{\lambda}
\]
yields
\[
b_j = B \frac{k}{k} + \frac{1}{2} \log_2 \left( \frac{\sigma_j^2}{\prod_{i=1}^{k} \sigma_i^2} \right)^{1/k} + \frac{1}{2} \log_2 \left( \frac{h_j}{\prod_{i=1}^{k} h_i} \right)^{1/k}
\]
for \( j = 1, \ldots, k \).
Note that \( b \) is optimal since \((\lambda, b)\) is unique and \( \frac{\partial}{\partial x} g(x) = 1 \) for all \( j \). □

Remarks:
- One can show that \( D(b) = f(b) \) is a strictly convex function. This implies (without using the Lagrange multiplier method) that the CMP has a unique solution.
- A shorter, but less intuitive proof of the bit allocation formula can be given using the convexity of the logarithm and Jensen’s inequality.
Minimum distortion:

\[ D_{\text{opt}} = \sum_{i=1}^{k} h_i \sigma_i^2 2^{-2b_i} \]

\[ = \sum_{i=1}^{k} h_i \sigma_i^2 2^{-2 \left( \bar{b} + \log_2 \frac{\sigma_i^2}{\rho^2} \right) - \log_2 \frac{\bar{b}}{\rho^2}} \]

\[ = \sum_{i=1}^{k} h_i \sigma_i^2 2^{-\log_2 \frac{\sigma_i^2}{\rho^2} 2^{-2\bar{b}}} \]

\[ = \sum_{i=1}^{k} \rho^2 H 2^{-2\bar{b}} \]

\[ = kH \rho^2 2^{-2\bar{b}} \]

**Note:** The same distortion is achieved if \( k \) random variables with common “shape factor” \( H \) and variance \( \rho^2 \) are optimally scalar quantized at a rate \( \bar{b} \) bits each.

---

**Transform coding: linear algebra review**

**Definitions:** Let \( A \) be a \( k \times k \) matrix of real elements.

- If \( A \) is symmetric and \( x^t A x \geq 0 \) for all \( x \in \mathbb{R}^k \), then \( A \) is called **nonnegative definite**.
- If \( A \) is symmetric and \( x^t A x > 0 \) for all \( x \neq 0 \), then \( A \) is called **positive definite**.
- \( A \) is **orthogonal** if
  
  \[ A^t = A^{-1} \quad \text{i.e.} \quad AA^t = I \]
  
  where \( I \) is the \( k \times k \) identity matrix.

**Fact 1:** If \( A \) is symmetric and nonnegative (positive) definite, then all of its eigenvalues are nonnegative (positive).

**Fact 2:** If \( A \) is symmetric and nonnegative definite, then it has \( k \) eigenvalues (counting multiplicities) and corresponding \( k \) mutually orthogonal eigenvectors.

**Fact 3:** If \( A = \{a_{ij}\} \) is a \( k \times k \) matrix with eigenvalues \( \lambda_1, \ldots, \lambda_k \) (counting multiplicities), then

\[ \text{trace}(A) \triangleq \sum_{i=1}^{k} a_{ii} = \sum_{i=1}^{k} \lambda_i \]

and

\[ \det(A) = \prod_{i=1}^{k} \lambda_i \]
**Fact 4** If $A$ is orthogonal, then it is norm-preserving, i.e.
$$
\|Ax\| = \|x\| \quad \text{for all } x \in \mathbb{R}^k
$$
(Here $\|y\| = \left( \sum_{i=1}^{k} y_i^2 \right)^{1/2}$ for any $y = (y_1, \ldots, y_k)^t$.)

**Fact 5** If $A$ is orthogonal, then $|\det(A)| = 1$.

**Fact 6** Let
$$X = (X_1, \ldots, X_k)^t$$
be a vector of real random variables having finite variance. Then the $k \times k$ autocorrelation matrix $R_X = \{E(X_iX_j)\}$ is symmetric and nonnegative definite.

*Proof:* We proved this when we derived the optimal linear predictor coefficients.

**Fact 7** If $X = (X_1, \ldots, X_k)^t$, $A$ is a $k \times k$ matrix, and $Y = AX$, then
$$R_Y = AR_XA^t$$

*Proof:* We have
$$R_Y = E \left[ \begin{array}{cccc} Y_1Y_1 & Y_1Y_2 & \cdots & Y_1Y_k \\ Y_2Y_1 & Y_2Y_2 & \cdots & Y_2Y_k \\ \vdots & \vdots & \ddots & \vdots \\ Y_kY_1 & Y_kY_2 & \cdots & Y_kY_k \end{array} \right] = E[YY^t] = E[AX(AX)^t] = A E[XX^t] A^t = AR_XA^t \square$$

**Fact 8** If $Y = AX$, then
$$\det(R_Y) = \det(A)^2 \det(R_X)$$

If $A$ is orthogonal, then
$$\det(R_Y) = \det(R_X)$$

*Proof:* Since $R_Y = AR_XA^t$,
$$\det(R_Y) = \det(A) \det(R_X) \det(A^t) = \det(A)^2 \det(R_X)$$

If $A$ is orthogonal, then $\det(A)^2 = 1$ by Fact 5, so $\det(R_Y) = \det(R_X)$. \square

---

**Transform coding with scalar quantization**

- $A$ is a $k \times k$ orthogonal matrix
- $X = (X_1, \ldots, X_k)^t$, $AX = Y = (Y_1, \ldots, Y_k)^t$.
  - The $Y_i$ are called the *transform coefficients*.
- Each $Q_i$ is an $N_i$-level scalar quantizer.
- $\hat{Y} = (Q_1(Y_1), \ldots, Q_k(Y_k))^t$, $A^{-1}\hat{Y} = \hat{X} = (\hat{X}_1, \ldots, \hat{X}_k)^t$
The end-to-end MSE distortion in transform coding is

\[ D_{tc} = \sum_{i=1}^{k} E[(X_i - \hat{X}_i)^2] = E[\|X - \hat{X}\|^2] \]

**Proposition 1**

In transform coding

\[ \|X - \hat{X}\| = \|Y - \hat{Y}\| \]

so that

\[ D_{tc} = E[\|Y - \hat{Y}\|^2] = \sum_{i=1}^{k} E[(Y_i - Q_i(Y_i))^2] \]

*Note:* The overall mean square reconstruction error is equal to the MSE distortion incurred by quantizing the transform coefficients.

---

**Proof of Proposition:**

\[
\|Y - \hat{Y}\| = \|AX - AA^{-1}\hat{Y}\| \\
= \|A(X - \hat{X})\| \quad \text{(since } A^{-1}\hat{Y} = \hat{X}) \\
= \|X - \hat{X}\| \quad \text{(by Fact 4)}
\]

Thus

\[ D_{tc} = E[\|X - \hat{X}\|^2] = E[\|Y - \hat{Y}\|^2] \]

□

What is a good choice for the transform \( A \)?

---

**Karhunen-Loeve (KL) transform**

- Order the eigenvalues \(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0\)
  - \(R_X\) (Fact 1) as
  - and let \(u_1, \ldots, u_k\) be the corresponding orthogonal eigenvectors (Fact 2).
- Normalize the \(u_i\) to unit length: \(\|u_i\| = 1, i = 1, \ldots, k\).
- Define \(U = [u_1 \ u_2 \ \ldots \ u_k]\) and let
  - \(T = U^t\)
- \(T\) is called the *Karhunen-Loeve transform (KLT)* matrix for \(X\).

*Note:* \(T\) is *orthogonal* since \(TT^t = U^tU = I\).

---

**Remarks:**

- In general \(T\) is not unique (different orderings of the eigenvectors give rise to different \(T\)’s).
- In statistics, the transform coefficients \((Y_1, \ldots, Y_k)^t = TX\) are called the *principal components* of \(X\).
- Since \(Y = TX = U^tX\), we have
  - \(Y_i = u_i^tX\)
  - and
  - \(X = IX = UU^tX = U(Y_1, \ldots, Y_k)^t = \sum_{i=1}^{k} Y_iu_i\)

Thus the transform coefficient \(Y_i\) represents the \(i\)th coordinate of \(X\) with respect to the orthogonal basis \(u_1, \ldots, u_k\).
Proposition 2

The transformation \( Y = TX \) decorrelates \( X \), i.e., \( Y = (Y_1, \ldots, Y_k)^t \) satisfies

\[
E(Y_i Y_j) = 0 \quad \text{for all } i \neq j
\]

Proof: Since \( Y = U^t X \), from Fact 7 we have

\[
R_Y = U^t R_X U
\]

But \( R_X u_i = \lambda_i u_i \) so

\[
R_X U = [R_X u_1 \ R_X u_2 \ldots R_X u_k] = [\lambda_1 u_1 \ \lambda_2 u_2 \ldots \lambda_k u_k]
\]

Hence

\[
R_Y = U^t R_X U = \begin{bmatrix}
  u_1^t \\
  u_2^t \\
  \vdots \\
  u_k^t
\end{bmatrix}
\begin{bmatrix}
  \lambda_1 & 0 & \ldots & 0 \\
  0 & \lambda_2 & \ldots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \ldots & \lambda_k
\end{bmatrix}
\]

Since \( R_Y = \{E(Y_i Y_j)\} \), this means

\[
E(Y_i Y_j) = \begin{cases}
  \lambda_i & \text{if } i = j \\
  0 & \text{if } i \neq j
\end{cases}
\]

Note:

- \( E(Y_i^2) = \lambda_i \). Thus the second moment of \( Y_i \) is the \( i \)th eigenvalue of \( R_X \).
- The proof demonstrates that any \( k \times k \) nonnegative definite \( R \) with eigenvalues \( \lambda_1, \ldots, \lambda_k \) can be written as

\[
R = U \Lambda U^t
\]

where \( U \) is orthogonal and \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k) \) is diagonal.

(This is the spectral theorem and it holds more generally, e.g., for symmetric matrices.)

We will show that the KLT is “optimal” under certain conditions. We need some auxiliary results.

Lemma 4 (Arithmetic-geometric mean inequality)

For any sequence of \( k \) nonnegative real numbers \( b_1, \ldots, b_k \)

\[
\left( \prod_{i=1}^k b_i \right)^{1/k} \leq \frac{1}{k} \sum_{i=1}^k b_i
\]

where equality holds if and only if \( b_i = b, \ i = 1, \ldots, k \).

Proof: If \( b_i = 0 \) for some \( i \), the inequality is trivial. Assuming \( b_i > 0 \) for all \( i \), the inequality is equivalent to

\[
\frac{1}{k} \sum_{i=1}^k \ln b_i \leq \ln \left( \frac{1}{k} \sum_{i=1}^k b_i \right)
\]

which follows from Jensen’s inequality and the convexity of the logarithm. Condition for equality also follows from Jensen.

Lemma 5 (Hadamard’s Inequality)

Let \( R \) be a nonnegative definite matrix with diagonal elements \( r_{ii} \), \( i = 1, \ldots, k \). Then

\[
\det(R) \leq \prod_{i=1}^k r_{ii}
\]

If \( R \) is positive definite, then equality holds if and only if \( R \) is diagonal.
Proof of Lemma 5: If $\mathbf{R}$ is not positive definite, then $\det(\mathbf{R}) = 0$. Since $r_{ii} \geq 0$ for all $i$, the inequality holds.
Assume $\mathbf{R}$ is positive definite. Then $r_{ii} > 0$ for all $i$.
Let $s_i = \frac{1}{\sqrt{r_{ii}}}$ and define
$$\mathbf{S} = \text{diag}(s_1, s_2, \ldots, s_k)$$
Consider
$$\mathbf{R} = \mathbf{S} \mathbf{R} \mathbf{S}^T$$
and note that
$$\det(\mathbf{R}) = \det(\mathbf{S})^2 \det(\mathbf{R}) = \left(\prod_{i=1}^{k} \frac{1}{\sqrt{r_{ii}}}\right) \det(\mathbf{R})$$
Thus
$$\det(\mathbf{R}) \leq \prod_{i=1}^{k} r_{ii} \iff \det(\mathbf{R}) \leq 1$$

Condition for equality:
- If $\mathbf{R}$ is diagonal, then $\det(\mathbf{R}) = \prod_{i=1}^{k} r_{ii}$.
- Conversely, if equality holds in $(\ast)$ and $(\ast\ast)$, then $\mu_i = \mu$ for all $i = 1, \ldots, k$.
Since $\mathbf{R}$ is symmetric and positive definite, by the spectral theorem
$$\mathbf{R} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$$
where $\mathbf{\Lambda} = \text{diag}(\mu_1, \ldots, \mu_k)$ and $\mathbf{U}$ is orthogonal.
If $\mu_i = \mu$ for all $i$, then $\mathbf{\Lambda} = \mu \mathbf{I}$ and
$$\mathbf{R} = \mathbf{U}(\mu \mathbf{I}) \mathbf{U}^T = \mu \mathbf{U} \mathbf{U}^T = \mu \mathbf{I}$$
Thus
$$\mathbf{R} = \mathbf{S}^{-1} \mathbf{R} \mathbf{S}^{-1} = \mathbf{S}^{-1} (\mu \mathbf{I}) \mathbf{S}^{-1} = \mu \mathbf{S}^{-2}$$
Since
$$\mathbf{S}^{-2} = \text{diag}(s_1^{-2}, \ldots, s_k^{-2})$$
we obtain that $\mathbf{R} = \mu \mathbf{S}^{-2}$ is diagonal. □

Note that $\mathbf{R} = \mathbf{S} \mathbf{R} \mathbf{S}^T$ with $\mathbf{S} = \text{diag}(s_1, s_2, \ldots, s_k)$ implies
$$\hat{r}_{ij} = s_i r_{ij} s_j$$
so that $\hat{r}_{ii} = \frac{r_{ii}}{(\sqrt{r_{ii}})(\sqrt{r_{ii}})} = 1$, $i = 1, \ldots, k$.
$\mathbf{R}$ is clearly symmetric and positive definite. Let $\mu_1, \ldots, \mu_k$ be the (positive) eigenvalues of $\mathbf{R}$. Then from Fact 3 and the arithmetic-geometric mean inequality
$$\det(\mathbf{R})^{1/k} = \left(\prod_{i=1}^{k} \mu_i\right)^{1/k} \leq \frac{1}{k} \sum_{i=1}^{k} \mu_i = \frac{1}{k} \text{trace}(\mathbf{R}) = 1$$
This is equivalent to
$$\det(\mathbf{R}) \leq 1$$
i.e.,
$$\det(\mathbf{R}) \leq \prod_{i=1}^{k} r_{ii}$$

High-resolution optimality of the KLT
Recall that
$$D_{\text{wc}} = E[\|\mathbf{X} - \hat{\mathbf{X}}\|^2] = E[\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2] = \sum_{i=1}^{k} E[(Y_i - Q_i(Y_i))^2]$$
where $\mathbf{Y} = \mathbf{A} \mathbf{X}$ and $\mathbf{A}$ is orthogonal.
Assumptions:
- $\mathbf{X} = (X_1, \ldots, X_k)^T$ is a Gaussian random vector having zero mean such that $\mathbf{R} \mathbf{X}$ is nonsingular.
- The high-resolution approximation to the optimal scalar quantizer distortion as valid with equality.
- Each $Q_i$ is the optimal $b_i$-bit quantizer for $Y_i$.
- The $b_i$ are optimally allocated for the constraint $\sum_{i=1}^{k} b_i \leq B$, where the $b_i$ need not be positive integers.
Key observation: If $Y = AX$ and $X$ is Gaussian, then each $Y_i$ is Gaussian since $Y_i = \sum_{j=1}^{k} a_{ij} X_j$. Thus all the $Y_i$ have the same $h_i = h_y$ coefficient in the high-resolution formula:

$$E[(Y_i - Q_i(Y_i))^2] = \sigma_i^2 h_y 2^{-2b_i}$$

where $\sigma_i^2 = E(Y_i^2)$ and

$$h_y = \frac{1}{12} \left( \int_{-\infty}^{\infty} \left( \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \right)^{1/3} dx \right)^3$$

Performance gain:

- **Transform coding distortion:** With the KLT and optimal bit-allocation

$$D_{tc} = kh_y 2^{-2b} \det(R_X)^{1/k} = kh_y 2^{-2b} \left( \prod_{i=1}^{k} \lambda_i \right)^{1/k}$$

where $\lambda_1, \ldots, \lambda_k$ are the eigenvalues of $R_X$.

- **Scalar quantization without transform coding:** Suppose equal variance $E(X_i^2) = \sigma_X^2$ for all $i$. If each $X_i$ is independently and optimally scalar quantized, the optimal bit allocation is

$$b_i = \bar{b} = \frac{B}{k} \quad i = 1, \ldots, k.$$  

The resulting MSE is

$$D_{PCM} = \sum_{i=1}^{k} E[(X_i - Q_i(X_i))^2] = kh_y \sigma_X^2 2^{-2b}$$

Gain of transform coding:

- Since $E(X_i^2) = \sigma_X^2$ for all $i$

$$\sigma_X^2 = \frac{1}{k} \sum_{i=1}^{k} E(X_i^2) = \frac{1}{k} \text{trace}(R_X)$$

- Thus

$$\frac{D_{PCM}}{D_{tc}} = \left( \frac{\sigma_X^2}{\prod_{i=1}^{k} \lambda_i} \right)^{1/k} = \frac{1}{k} \sum_{i=1}^{k} \lambda_i \left( \prod_{i=1}^{k} \lambda_i \right)^{1/k} \geq 1$$

Equality holds if and only if all the $\lambda_i$ are equal, which happens if and only if $R_X = \text{diag}(\sigma_X^2, \ldots, \sigma_X^2)$.

We obtain that the gain is always greater than 1 unless the $X_i$ are independent.

- The more uneven the distribution of the $\lambda_i$, the more can be gained by transform coding.
The Gaussian source assumption is essential in the KLT optimality proof and the performance gain calculation.

The assumptions of high-resolution quantization and non-integer bit rates can be dropped. It can be shown without any further conditions that the KLT is the optimal transform for Gaussian random vectors if optimal scalar quantization and bit allocation are used.

Advantages of DCT

- Fixed (signal-independent).
- Easy to compute (using FFT).
- Very good energy compaction for highly correlated data.
- Good approximation to KLT for Markov sources. Often achieves performance close to that of the KLT.

Transform coding in practice

- The KLT depends on the source autocorrelation matrix $R_X$. It has to be recalculated based on estimates of $R_X$ if the source statistics changes, which is computationally expensive.
- The most popular fixed (source-independent) transform is the discrete cosine transform (DCT) given by the transform matrix $T = \{t_{mn}\}_{m,n=0}^{k-1}$:

$$
t_{mn} = \begin{cases} 
\sqrt{\frac{2}{k}} \cos\left(\frac{\pi}{k} m \left(n + \frac{1}{2}\right)\right) & m = 1, \ldots, k - 1, \ n = 0, \ldots, k - 1 \\
\sqrt{\frac{1}{k}} \cos\left(\frac{\pi}{k} m \left(n + \frac{1}{2}\right)\right) & m = 0, \ n = 0, \ldots, k - 1
\end{cases}
$$

Can be shown:

- DCT is an orthogonal transform
- The DCT of a vector $x = (x_0, \ldots, x_{k-1})$ can be obtained using the FFT:
  - Define the mirrored sequence $y = (x_0, \ldots, x_{k-1}, x_{k-1}, \ldots, x_0)$
  - Compute the FFT of $y$. The DCT coefficients of $x$ are obtained by scaling the first $k$ FFT coefficients of $y$.
  - Complexity reduced from $O(k^2)$ to $O(k \log k)$. 
Transform coding of images

- We defined transform coding for “one dimensional” signals, but images are two-dimensional.
- Let \( X = \{X_{i,j}\}_{i,j=0}^{k-1} \) denote a block (matrix) of pixels:
  \[
  X = \begin{bmatrix}
  X_{0,0} & X_{0,1} & \ldots & X_{0,k-1} \\
  X_{1,0} & X_{1,1} & \ldots & X_{1,k-1} \\
  \vdots & \vdots & \ddots & \vdots \\
  X_{k-1,0} & X_{k-1,1} & \ldots & X_{k-1,k-1}
  \end{bmatrix}
  \]
  (Can be a sub-block of a larger image.)
- Let \( A \) be a \( k \times k \) orthogonal matrix and define the 2-D separable transform by
  \[
  Y = AXA^t = A(XA^t)
  \]
  Interpretation: First the rows of \( X \) are transformed by \( A \), then the columns of the resulting matrix are transformed by \( A \).
- The inverse transform for
  \[
  Y = AXA^t
  \]
  is given by
  \[
  X = A^tYA
  \]
- Let \( A^t = [a_0 \ a_1 \ \ldots \ a_{k-1}] \) (\( a_i^t \) is the \( i \)th row of \( A \)). It is easy to show that
  \[
  X = \sum_{i,j=0}^{k-1} Y_{i,j}B_{ij}
  \]
  where the \( k \times k \) matrix \( B_{i,j} \) is given by \( B_{i,j} = a_ia_j^t \).
- The \( B_{i,j} \) are called basis functions (matrices) for the 2-D separable transform \( A \).
- It can be shown that the \( B_{i,j} \) form an orthonormal basis for \( M_k(\mathbb{R}) \), the vector space of \( k \times k \) real matrices.

Example: The 64 basis functions for the 8 \( \times \) 8 2-D DCT
JPEG (Joint Photographic Experts Group) standard

- DCT transform coding on $8 \times 8$ pixel blocks.
- Transform coefficients are quantized using uniform quantizers; instead of explicit bit allocation, the step sizes are varied.
- Quantizer output is further compressed by lossless coding:
  - Runlength coding of quantized coefficients (efficient since many coefficients become zero after quantization);
  - Huffman coding of runs of zeros and quantizer output indices.

For color images, RGB $\rightarrow$ YCbCr color space conversion is applied first. The luminance (Y) and chrominance (Cb and Cr) components are separately encoded. The chrominance components are usually subsampled by a factor of 2.

**Performance (color photos):**
- fair - good : 0.25 bits/pixel (bpp)
- good - very good : 0.5 bpp
- excellent: 1.5 bpp
- indistinguishable: above 2.5 bpp

Compare with original $3 \times 8 = 24$ bpp.

**Note:** At 0.25 bpp, the blocking effects are clearly visible.
Subband coding

- Pass signal through a filter bank and quantize the outputs (subbands) separately. Each subband is down-sampled according to the ratio of the bandwidth of the filter output and input.
- Use bit allocation among subband quantizers to minimize overall distortion.
- In image coding practice, often a cascade of lowpass-highpass filter pairs is used.

Data Compression and Source Coding V: Transform Coding

Application to image compression:
- Instead of uniform subbands, first break up signal into high and low frequency subbands using a lowpass-highpass filter pair. Divide only the low frequency subbands into further high and low frequency subbands.
- Reconstruction reverses process. Begin with the lowest frequency subband and successively add back in higher frequencies.
- For images, do this separately on rows and columns.
- “Multiresolution”: natural means of reconstructing larger and higher resolution images working from lower to higher frequencies.
- Can use e.g. quadrature mirror filters (QMF) of wavelet decomposition (JPEG 2000).
Single pass of subband decomposition of image using wavelet transform.

*Note:* The highpass components are almost invisible because most of the signal energy is concentrated in the low-low image subband.

⇒ Energy compaction.

### DCT vs. wavelet-based coding

**JPEG:** 0.25 bpp  
**JPEG2000:** 0.25 bpp

*Note:* JPEG blocking effects are clearly visible, but some “high frequency” details show better on the JPEG picture.

**JPEG:** 0.25 bpp  
**JPEG2000:** 0.25 bpp

*Note:* JPEG2000 is clearly superior here.