Lattice vector quantization

Let \( \{u_1, \ldots, u_n\} \) be a set of linearly independent vectors in \( \mathbb{R}^n \).

Then the lattice \( \Lambda \) generated by \( \{u_1, \ldots, u_n\} \) is a set of all points of the form

\[
y = \sum_{i=1}^{n} c_i u_i, \quad \text{where} \quad c_i \quad \text{are integers.}
\]

The vectors \( \{u_i\} \) is a basis for \( n \) dimensional lattice.

The matrix \( U = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \) is a generator matrix of the lattice.

Any vector of the lattice \( y = cU, \quad c = (c_1, \ldots, c_n) \).
Lattice vector quantization

The hexagonal lattice $A_2$ has the generator matrix

$$U = \begin{pmatrix} 1 & 0 \\ 1/2 & \sqrt{3}/2 \end{pmatrix}.$$ 

Any vector $y = (y_1, y_2)$ of this lattice can be written

$$y_1 = c_1 + c_2 / 2$$
$$y_2 = c_2 \sqrt{3} / 2.$$ 

Lattice quantizer is a vector quantizer whose codebook is a lattice.

The approximating vectors are centroids of the Voronoi polyhedrons which are of the same size and shape, t.i. congruent.
Lattice vector quantization

Voronoi cells of the hexagonal lattice
Lattice vector quantization

\( A_2 \) is a union of the scaled rectangular lattice and its translate:

Any vector of \( S^2 \) has the form \( y_1 = c_1 + k, k\sqrt{3} \) \hspace{1cm} (3.1)

and any vector of its translate has the form

\[ y_2 = (c_1 + k + 1/2, k\sqrt{3} + \sqrt{3}/2) = y_1 + (1/2, \sqrt{3}/2). \] \hspace{1cm} (3.2)

Let \( x = (-5.6, 0.82) \). In (3.1) it is quantized to \((-6.0, 0)\).

In (3.2) the approximation is \((-5.5, \sqrt{3}/2)\).

The best approximation is \((-5.5, \sqrt{3}/2)\). The corresponding error \( \|x - y\|^2 / 2 = 0.0058 \). The output: \( c_1 = -6, c_2 = 1 \).
Lattice quantizer

The special class of lattices are lattices based on linear codes.

Let $C$ be an $(n, k)$ binary linear code. Then the lattice $\Lambda(C)$ is defined 

$$\Lambda(C) = \{y \in \mathbb{Z}^n | y \equiv c (\text{mod } 2) \text{ for some } c \in C\}.$$ 

Assume that the generator matrix is systematic and given as 

$$G = (I | B).$$

Then $U = \begin{pmatrix} I & B \\ 0 & 2I \end{pmatrix}$ and 

$$\Lambda(C) = \bigcup_{i=0}^{2^k-1} (c_i + 2\mathbb{Z}^n).$$
Lattice quantizer. Quantization procedure

For each of $2^k$ cosets of $2\mathbb{Z}^n$ do the following:

- Subtract the corresponding codeword from the input vector $d_i = x - c_i$
- Scalar quantize each component of $d_i$ with step 2 and obtain the quantized vector $q_i$
- Compute the quantization error $\|x - y_i\|^2$, $y_i = 2q_i + c_i$
- Find the closest pair $(q_i, c_i)$ minimizing the error.
- Keep or transmit $(q_i, i)$. 
Lattice quantizer. Dequantization procedure.

• Reconstruct the approximating vector $\hat{d}_i$

• Add the corresponding codeword to the approximating vector $y_i = \hat{d}_i + c_i$. 
Lattice quantizer. Example.

Let $C$ be a $(3,2)$ linear block code with the generator matrix

$$G = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}. $$

The generator matrix of the lattice based on $C$ is

$$U = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}. $$

$x = (0.4,1.2,3.7)$

$c_0 = (0,0,0)$ $q_0 = (0,1,2)$ $d_0 = (0,2,4)$ $y_0 = (0,2,4)$ $D_0 = 0.297$

$c_1 = (0,1,1)$ $q_1 = (0,0,1)$ $d_1 = (0,0,2)$ $y_1 = (0,1,3)$ $D_1 = 0.230$

$c_2 = (1,0,1)$ $q_2 = (0,1,1)$ $d_2 = (0,2,2)$ $y_2 = (1,2,3)$ $D_2 = 0.497$

$c_3 = (1,1,0)$ $q_3 = (0,0,2)$ $d_3 = (0,0,4)$ $y_3 = (1,1,4)$ $D_3 = 0.163$

$i = 3$
Lattice quantizer

• The quantization procedure can be interpreted as a method for choosing the best sequence of $n$ scalar values among $2^k$ allowed sequences, where each approximating value is generated by one of the two scalar quantizers.

• The first scalar quantizer has the approximating values: ..., -4, -2, 0, 2, 4, .... The second quantizer has the approximating values ..., -3, -1, 1, 3, ...

• Finding the quantized vector = exhaustive search among $2^k$ codewords. The computational complexity can be reduced by using code trellis and the Viterbi decoding algorithm.
Elements of rate-distortion theory.

Rate-distortion function.

Each quantization procedure is characterized by the average distortion and by quantization rate.

The goal of compression system design is to optimize the rate-distortion tradeoff. In order to compare different quantizers the rate-distortion function $R(D)$ is introduced.

We say that for a given source a quantizer with rate-distortion function $R_1(D)$ is better than the other quantizer with $R_2(D)$ for $D = D_0$ if $R_1(D_0) \leq R_2(D_0)$.

The theoretical limit of rate-distortion functions is achievable rate-distortion function $H(D)$. 
Rate-distortion function for memoryless source

Theorem 3.1 The rate-distortion function $H(D)$ for $N(0, \sigma^2)$ source with squared error distortion is

$$H(D) = \begin{cases} \frac{1}{2} \log_2(\sigma^2 / D), & D \leq \sigma^2, \\ 0, & D > \sigma^2 \end{cases} \quad (3.3)$$

where $N(0, \sigma^2)$ denotes the Gaussian variable with zero mean value and variance $\sigma^2$.

In general case $H(D)$

• Is non-increasing and convex downwards function of $D$.

• There exists some value $D_0$ such that $H(D) = 0$, for all $D \geq D_0$. 
Rate-distortion function for memoryless source

Fig. 3.1
Rate-distortion function for source with memory

Consider a source which generates the stationary random Gaussian process of discrete time.

That is for any $n = 1,2,...$ a random vector $\mathbf{X} = (X_1,...,X_n)$ at the output of the source is Gaussian vector with covariance matrix $\Lambda_n$ and vector of average values $\mathbf{m}$.

Its pdf has the form

$$f_n(x) = \frac{1}{(2\pi)^{n/2}|\Lambda_n|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{m})\Lambda_n^{-1}(\mathbf{x} - \mathbf{m})^T\right\},$$

$$\Lambda_n = E\{(\mathbf{X} - \mathbf{m})^T(\mathbf{X} - \mathbf{m})\} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \ldots & \lambda_{nn} \end{pmatrix}.$$
Rate-distortion function for source with memory

\[ \lambda_{ij} = E \{(X_i - m_i)(X_j - m_j)\} \] is the covariance moment of \( X_i \) and \( X_j \). \( \lambda_{ii} \) is the variance of \( X_i \).

The property of stationarity means that the dimensional pdfs of vectors \( \mathbf{X} = (X_1, \ldots, X_n) \) and \( \mathbf{X}_j = (X_{j+1}, \ldots, X_{j+n}) \) are identical. It follows from stationarity that \( \lambda_{ij} = \lambda_{ji} = \lambda_{\mid i-j \mid} \) and

\[
\Lambda_n = \begin{pmatrix}
\lambda_0 & \lambda_1 & \ldots & \lambda_{n-1} \\
\ldots & \ldots & \ldots & \ldots \\
\lambda_{n-1} & \lambda_{n-2} & \ldots & \lambda_0
\end{pmatrix}
\]
is Toeplitz’s matrix.
Rate-distortion function for source with memory

Let \(|i - j| = \tau\). Assume that \(m = 0\) and that \(\lim_{\tau \to \infty} \lambda_\tau = 0\).

The Fourier series expansion

\[
\sum_{\tau = -\infty}^{\infty} \lambda_\tau e^{-j2\pi f \tau} = N(f), \quad -1/2 \leq f \leq 1/2
\]

is called power spectral density of the random process generated by a stationary source. It characterizes how the power of the process is distributed over frequencies.

\[
\lambda_\tau = \int_{-1/2}^{1/2} N(f)e^{j2\pi f \tau} df.
\]
Rate-distortion function for source with memory
The main properties of the power spectral density are:

- $N(f)$ is a real function
- $N(f) = 2 \sum_{\tau=1}^{\infty} \lambda_{\tau} \cos(2\pi f \tau) + \lambda_0$ since $\lambda_{\tau} = \lambda_{-\tau}$
- It follows from the previous property that $N(f)$ is even function and thereby $\lambda_{\tau} = \int_{-1/2}^{1/2} N(f) \cos(2\pi f \tau) df$
- $N(f) \geq 0$ for all $f \in [-1/2,1/2]$
- $\lambda_0 = \int_{-1/2}^{1/2} N(f) df$ is the variance of the random process with power spectral density $N(f)$. 

Rate-distortion function for source with memory

Theorem 3.2 The rate-distortion function $H(D)$ with squared error distortion for the discrete time stationary random Gaussian process with bounded and integrable spectral density $N(f)$ is computed as

$$H(D) = \frac{1}{2} \int_{-1/2}^{1/2} \log_2 \left\{ \max \left\{ 1, \frac{N(f)}{\theta} \right\} \right\} df,$$

(3.4)

$$\int_{-1/2}^{1/2} \min \{ \theta, N(f) \} df = D.$$
Rate-distortion function for source with memory

Interpretation (3.4) by using “water-filling”.

\[ N(f) \]

\[ f_{-1} \quad -f_1 \quad -f_2 \quad -f_3 \quad 0 \quad f_3 \quad f_2 \quad f_1 \quad \frac{1}{2} \quad f \]

\[ \theta \]
Rate-distortion function for source with memory

\[ H(D) \]

\[ \rho = 0 \]
\[ \rho = 0.5 \]
\[ \rho = 0.99 \]
Coding theorems

Theorem 3.3 For discrete time stationary continuous source with rate-distortion function $H(D)$ with respect to the squared error $d(x, y) = (x - y)^2$ there exists such $n_0$ that for all $n > n_0$ and for $\delta_1 > 0$ and $\delta_2 > 0$ there exists an $(R, D_n)$ code of codelength $n$ with coderate $R \leq H(D) + \delta_1$ for which the MSE $D_n$ is less than or equal to $D + \delta_2$.

Theorem 3.4 (Converse of Th.3.3). For the source from Theorem 3.3 there does not exist a code for which simultaneously the MSE would be less than or equal to $D$ and $R < H(D)$. 
Comparison of quantization procedures

For the uniform scalar quantizer if the quantization step $\delta$ is small enough we can assume that pdf is constant inside each cell, that is $P(y_i) \approx \delta f(y_i)$. Then we obtain

$$D \approx \sum_j \int (x - y_j)^2 f(y_j) dx \approx \sum_j \frac{P(y_j)}{\delta} \int_{y_j - \delta/2}^{y_j + \delta/2} (x - y_j)^2 dx = \frac{\delta^2}{12}.$$  

The rate of such a quantizer:  

$$R(D) = h(X) - \frac{1}{2} \log_2 (12D),$$

$$h(X) = -\int_X f(x) \log_2 f(x) dx$$

is the differential entropy of $X$.

For Gaussian memoryless source:  

$$H(D) = h(X) - \frac{1}{2} \log_2 (2\pi eD)$$

$$R(D) - H(D) \leq \frac{1}{2} \log_2 \frac{\pi e}{6} \approx 0.2546$$
Comparison of quantization procedures

Assume that number of cells of the lattice quantizer is large and pdf is constant within each cell \( S_i, \ i = 1, \ldots, M \).

\[
D_n \approx \frac{1}{n} \int \frac{\|x\|^2}{\text{Vol}(S)} \ dx
\]

\[
R \approx h(X) - \frac{1}{n} \log_2 \text{Vol}(S),
\]

\[
\text{Vol}(S) = \int_S dx \quad \text{is the volume of the Voronoi cell} \ S.
\]

\[
R(D) \approx H(D) + \frac{1}{2} \log_2 2\pi e G_n,
\]

\[
G_n = \frac{1}{n} \frac{\int_S \|x\|^2}{\text{Vol}(S)^{1+2/n}}
\]

is the normalized second moment of the Voronoi region.
Comparison of quantization procedures

The following bounds on \( G_n \) hold

\[
\frac{1}{(n+2)\pi} \Gamma\left(\frac{n}{2} + 1\right)^{2/n} \leq G_n \leq \frac{1}{n\pi} \Gamma\left(\frac{n}{2} + 1\right)^{2/n} \Gamma\left(1 + \frac{2}{n}\right),
\]

\( G_n = \frac{1}{(n+2)\pi} \Gamma\left(\frac{n}{2} + 1\right)^{2/n} \) is the normalized second moment of the n-dimensional sphere.

If \( n \to \infty \) \( G_n \to \frac{1}{2\pi e} = 0.058550... \) and \( R(D) \approx H(D) \)

for Gaussian memoryless source.
Pdf of the generalized Gaussian distribution

\[ f(x) = \frac{\alpha \gamma(\alpha, \sigma)}{2\Gamma(1/\alpha)} \exp\{-(\gamma(\alpha, \sigma)|x - m|^\alpha}\}, \]

where \( m \) is the mathematical expectation, \( \sigma^2 \) is the variance, \( \alpha \) is a parameter,

\[ \gamma(\alpha, \sigma) = \sigma^{-1} \left[ \frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)} \right]. \]

\[ \Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt, \]

\( \Gamma(1) = 1, \Gamma(x + 1) = x\Gamma(x), \)

\( \Gamma(n + 1) = n!, \)

\( \Gamma(1/2) = \sqrt{\pi}. \)
Pdf of the generalized Gaussian distribution

\[ f(x) \]

\[ \alpha = 1.0 \]
\[ \alpha = 0.5 \]
\[ \alpha = 2.0 \]
Comparison of fixed-rate quantizers ($\alpha = 2.0$)
Comparison of fixed-rate quantizers ($\alpha = 0.5$)
Comparison variable-rate quantizers ($\alpha = 2.0$)
Comparison variable-rate quantizers ($\alpha = 0.5$)

\[
\log_{10} \approx \frac{1}{\alpha} \\
H(D), R(D)
\]
Comparison of variable-rate quantization procedures

Fixed-rate quantizers:

• **Nonuniform scalar quantizer** is always better than uniform scalar quantizer

• Using **scalar** nonuniform quantizer it is impossible to obtain coderate less than 1 bit/sample

• **Vector quantization reduces redundancy** $\max_D (R(D) - H(D))$ compared to scalar quantizer. Besides that VQ provides coderates less than 1 bit/sample

• VQ for the Gaussian stationary process with $n \to \infty$ might lead to the curve $R(D)$ lying below $H(D)$ in Fig.3.1
Comparison of quantization procedures

If a quantizer is not constrained to be a fixed-rate then its outputs can be lossless coded by a variable-length encoder.

It can be shown the redundancy of the variable-rate uniform quantizer when $\delta \to 0$ is

$$C \leq \frac{1}{2} \log_2 \frac{\pi e}{6} = 0.2546.$$  

- For rates greater than 2 bits/sample uniform quantizer (rounding off), optimal scalar quantizer (ECSQ) provide almost the same $R(D)$.

- For rates less than 2 bit/sample uniform quantizer (rounding off) has worse performance than ECSQ.

- For rates less than 2 bits/sample suboptimal scalar quantizers with extended zero zone and optimal uniform quantizer have performances rather close to the performances of ECSQ.
Comparison of quantization procedures

• The uniform (rounding off) quantizer followed by a variable-length coder can provide coderates less than 1 bit/sample.

• The performance of the entropy-constrained scalar quantizer is slightly superior to the performance of the uniform scalar quantizer followed by a variable-rate coder.

• Entropy-constrained scalar quantizer provides coderates less than 1 bit/sample.

• For memoryless source VQ reduces redundancy compared to scalar quantization.

• For source with memory only VQ allows to obtain $R(D)$ close to $H(D)$. 
## Characteristics of digital speech, audio, image, and video signals

<table>
<thead>
<tr>
<th>Speech and audio type</th>
<th>Frequency band</th>
<th>Sampling rate</th>
<th>Bits/sample</th>
<th>Uncompressed bit rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>200-3200Hz</td>
<td>8 kHz</td>
<td>16</td>
<td>128kb/s</td>
</tr>
<tr>
<td>CD audio</td>
<td>20-20000Hz</td>
<td>44.1kHz</td>
<td>16x2 channels</td>
<td>1.41 Mb/s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Still image</th>
<th>Pixels per frame</th>
<th>Bits/Pixel</th>
<th>Uncompressed bit rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAX</td>
<td>3120×2040</td>
<td>1</td>
<td>6.36 Mb</td>
</tr>
<tr>
<td>VGA</td>
<td>640×480</td>
<td>8</td>
<td>2.46 Mb</td>
</tr>
<tr>
<td>XVGA</td>
<td>1024×768</td>
<td>24</td>
<td>18.87 Mb</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video</th>
<th>Pixels per frame</th>
<th>Frame per Second</th>
<th>Bits/pixel</th>
<th>Uncompressed bit rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTSC</td>
<td>700×525</td>
<td>30</td>
<td>16</td>
<td>176.4 Mb/s</td>
</tr>
<tr>
<td>PAL</td>
<td>833×625</td>
<td>25</td>
<td>16</td>
<td>208.3 Mb/s</td>
</tr>
<tr>
<td>CIF</td>
<td>352×288</td>
<td>15</td>
<td>12</td>
<td>18.2 Mb/s</td>
</tr>
<tr>
<td>QCIF</td>
<td>176×144</td>
<td>10</td>
<td>12</td>
<td>3.0 Mb/s</td>
</tr>
<tr>
<td>HDTV</td>
<td>1280×720</td>
<td>60</td>
<td>12</td>
<td>622.9 Mb/s</td>
</tr>
<tr>
<td>HDTV</td>
<td>1920×1080</td>
<td>30</td>
<td>12</td>
<td>745.7 Mb/s</td>
</tr>
</tbody>
</table>