Scalar Quantization

Original purpose: to discretize “analog” values. For example

\[ x \rightarrow \left\lfloor x + \frac{1}{2} \right\rfloor \] (round to the nearest integer)

**Definition (Scalar quantizer)**

An \( N \)-point scalar quantizer is a mapping

\[ Q : \mathbb{R} \rightarrow C \]

where \( C = \{y_1, y_2, \ldots, y_N\} \subset \mathbb{R} \).

- \( C \) is called the codebook of \( Q \).
- \( y_1, \ldots, y_N \) are the output levels (or output points, reproduction points, quantization points, etc).
- We always assume indexing such that \( y_1 < y_2 < \cdots < y_N \).

**Encoder-Decoder structure**

- **Encoder**: \( E : \mathbb{R} \rightarrow \{1, \ldots, N\} \) defined by \( E(x) = j \iff x \in R_j \)
- **Decoder**: \( D : \{1, \ldots, N\} \rightarrow \{y_1, \ldots, y_N\} \) defined by \( D(j) = y_j \)
- Thus \( Q(x) = D(E(x)) \)

**Rate of \( Q \)**:

\[ R(Q) = \log_2 N \text{ bits/sample} \]

- \( R \) is the number of bits needed to represent the \( N \) indices with binary words of a fixed length (fixed-rate binary encoding)

The sets \( R_i = \{x : Q(x) = y_i\}, \ i = 1, \ldots, N \) are called the quantizer cells.

Clearly

\[ R_i \cap R_j = \emptyset \quad \text{if} \quad i \neq j \]

and

\[ \bigcup_{i=1}^{N} R_i = \mathbb{R} \]

so that \( R_1, \ldots, R_N \) form a partition of \( \mathbb{R} \).

**Note**: \( Q \) is completely described by \( C \) and \( \{R_1, \ldots, R_N\} \) since

\[ Q(x) = y_i \quad \text{if and only if} \quad x \in R_i \]
**Definition (Regular quantizer)**

A quantizer \( Q \) is called a **regular** quantizer if

(a) each \( R_i \) is an interval \((x_{i-1}, x_i]\) (or \([x_{i-1}, x_i)\) or \((x_{i-1}, x_i)\), etc).

Note: \( x_{i-1} = \pm \infty \) and/or \( x_i = \pm \infty \) are not excluded

(b) \( y_i \in R_i \) for all \( i = 1, \ldots, N \).

- \( x_0, x_1, \ldots, x_N \) are called the **decision levels**
- For a regular quantizer \( x_0 < y_1 < x_1 < y_2 < \cdots < y_N < x_N \).

**Example:** 4-level regular quantizer

\[
\begin{array}{ccccccc}
& y_1 & x_1 & y_2 & x_2 & y_3 & x_3 & y_4 \\
\hline
R_3 & & & & & & & \\
\end{array}
\]

**Example:** Infinite-level uniform quantizer

Define

\[
Q(x) = \left\lfloor x + \frac{1}{2} \right\rfloor
\]

- The output levels are \( y_i = i, i = 0, \pm 1, \pm 2, \ldots, N = \infty \).
- The codebook is \( C = \mathbb{Z} \), the set of all integers.
- The quantizer cells are determined as:

\[
Q(x) = i \iff i \leq x + \frac{1}{2} < i + 1 \iff i - \frac{1}{2} \leq x < i + \frac{1}{2}
\]

so

\[
R_i = \left[ i - \frac{1}{2}, i + \frac{1}{2} \right), \quad i = 0, \pm 1, \pm 2, \ldots
\]

**Quality of reproduction**

- **Distortion measure:** \( d(x, y) \) measures the distortion (or loss) resulting if the input \( x \) is reproduced as \( y \).
- Mathematically, we require \( d(x, y) \geq 0 \) for all \( x, y \in \mathbb{R} \).
- Instantaneous distortion of \( Q \) for input \( x \):

\[
d(x, Q(x))
\]

- A random variable (r.v.) \( X \) is often called a **random source**.

**Expected distortion of \( Q \):**

\[
D(Q) = Ed(X, Q(X))
\]

To be useful, the distortion measure should be

- tractable (easy to compute)
- meaningful for perception or application

Some popular choices:

- \( d(x, y) = (x - y)^2 \) (squared error distortion, ubiquitous)
- \( d(x, y) = |x - y|^r \) (\( r > 0 \); \( r = 1 \) is the most common choice)
Assume

- The source $X$ has probability density function (pdf) $f$ (short notation: $X \sim f$);
- $Q$ is a regular quantizer with output levels $y_1, \ldots, y_N$ and decision levels $x_0, \ldots, x_N$.

Then the expected distortion is

\[
Ed(X, Q(X)) = \int_{-\infty}^{\infty} d(x, Q(x))f(x) \, dx = \sum_{i=1}^{N} \int_{R_i} d(x, Q(x))f(x) \, dx = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} d(x, y_i)f(x) \, dx
\]

**Example:** $N$-level *uniform quantizer* over the interval $(a, b]$:

- Let $\Delta = (b - a)/N$ and for $i = 1, \ldots, N$,
  
  $R_i = (a + (i - 1)\Delta, a + i\Delta]$  
  
  $y_i = a + \Delta \left( i - \frac{1}{2} \right)$  

Assume $X$ is uniformly distributed over $[a, b]$:

\[
f(x) = \begin{cases} 
\frac{1}{b - a} & \text{if } a \leq x \leq b \\
0 & \text{otherwise}
\end{cases}
\]

Consider the squared error $d(x, y) = (x - y)^2$.

**Signal-to-noise ratio**

Assume $d(x, y) = (x - y)^2$. Then $D(Q)$ is called the *mean squared error* (MSE) of $Q$.

- $e = Q(X) - X$ is called the quantization error (noise) since $Q(X) = X + Q(X) - X = X + e$
- $D(Q) = E[(X - Q(X))^2] = E(e^2)$
- *Signal-to-noise ratio* (signal-to-quantization noise ratio):
  
  \[
  \text{SNR} = 10 \log_{10} \frac{E(X^2)}{D(Q)} \text{ dB}
  \]

We conclude

\[
D(Q) = \frac{\Delta^2}{12}
\]
Performance of quantizers

The performance of an $N$-level quantizer is measured by its distortion and rate:

\[
D(Q) = \text{Ed}(X, Q(X)) \\
R(Q) = \log_2 N
\]

Goal: Find optimal tradeoffs between distortion and rate; i.e., minimize one while constraining the other.

Note: distortion and rate are conflicting quantities

Optimality of scalar quantizers

Definition

Let $Q_N$ denote the family of all $N$-level quantizers. $Q^* \in Q_N$ is an optimal quantizer if

\[
\text{Ed}(X, Q^*(X)) = \min_{Q \in Q_N} \text{Ed}(X, Q(X))
\]

Remarks:

- $Q^*$ depends on the distribution of $X$ and on $d$.
- $Q^*$ is not necessarily unique (there may be more than one optimal $N$-level quantizers).
- Can be shown that $Q^*$ exists for all $N$ and “reasonable” $d$ if $\text{Ed}(X, y) < \infty$ for some $y \in \mathbb{R}$.

Theorem 1 (Nearest Neighbor Condition)

Consider all $N$-level scalar quantizers with codebook $\mathcal{C} = \{y_1, \ldots, y_N\}$. Among these, any quantizer with quantization regions satisfying

\[
R_i \subset \{x : d(x, y_i) \leq d(x, y_j), \quad j = 1, \ldots, N\} \quad i = 1, \ldots, N \quad (*)
\]

has minimum distortion.

Note: $Q$ with codebook $\mathcal{C}$ satisfies $(*)$ if and only if for any $x$,

\[
Q(x) = y_i \quad \text{implies} \quad d(x, y_i) \leq d(x, y_j) \quad \text{for all} \quad j
\]

Equivalently, for all $x$,

\[
d(x, Q(x)) = \min_{y_j \in \mathcal{C}} d(x, y_j)
\]

or

\[
Q(x) = \arg \min_{y_j \in \mathcal{C}} d(x, y_j)
\]

“Proof” of Theorem 1: Let $Q \in Q_N$ with codebook $\mathcal{C}$ satisfy $(*)$, and let $\hat{Q} \in Q_N$ have codebook $\mathcal{C}$ but arbitrary quantization regions.

Then for all $x$,

\[
d(x, \hat{Q}(x)) \geq \min_{y_j \in \mathcal{C}} d(x, y_j) = d(x, Q(x))
\]

Thus

\[
\text{Ed}(X, \hat{Q}(X)) \geq \text{Ed}(X, Q(X))
\]
The subsets
\[ \tilde{R}_i = \{ x : d(x, y_j) \leq d(x, y_i), \ j = 1, \ldots, N \} \ i = 1, \ldots, N \]
cover the real line (i.e., \( \bigcup_{i=1}^{N} \tilde{R}_i = \mathbb{R} \)), but they are not disjoint.

A simple tie-breaking rule giving an optimal partition is
\[
R_1 = \tilde{R}_1, \quad R_i = \tilde{R}_i \setminus \bigcup_{j=1}^{i-1} R_i \quad \text{for } i = 2, \ldots, N
\]

To uniquely define \( R_i \), we choose
\[
R_i = (x_{i-1}, x_i), \quad \text{for } i = 2, \ldots, N - 1
\]

Given \( C \), the optimal quantizer is regular. The decision levels are the midpoints between two neighboring output levels.

**Theorem 2 (Centroid Condition)**

Consider all \( N \)-level scalar quantizers with a given partition \( \{R_1, \ldots, R_N\} \). Among these, the quantizer \( Q \) with output levels
\[
y_i = \arg\min_{y \in \mathbb{R}} E[d(X, y)|X \in R_i], \quad i = 1, \ldots, N
\]
has minimum distortion.

- \( y_i \) is called the centroid of \( R_i \).
- The centroid \( y_i \) may not be unique.

**Special Case:** \( d(x, y) = |x - y|^r \) (includes MSE)

\[
|x - y_i|^r \leq |x - y_j|^r \iff |x - y_i| \leq |x - y_j|
\]

Thus
\[
\tilde{R}_i = \{ x : |x - y_i| \leq |x - y_j|, \ j = 1, \ldots, N \} \ i = 1, \ldots, N
\]

If \( y_1 < y_2 < \cdots < y_N \), then \( \tilde{R}_i = [x_{i-1}, x_i] \), where
\[
x_{i-1} = \frac{y_{i-1} + y_i}{2} \quad \text{and} \quad x_i = \frac{y_i + y_{i+1}}{2}, \quad i = 2, \ldots, N - 1
\]

Proof of Theorem 2: Let \( P_i = P(X \in R_i) \). For any \( \hat{Q} \) with partition \( \{\hat{R}_1, \ldots, \hat{R}_N\} \) and arbitrary output levels \( \{\hat{y}_1, \ldots, \hat{y}_N\} \),

\[
Ed(X, \hat{Q}(X)) = \sum_{i=1}^{N} E[d(X, \hat{Q}(X))|X \in R_i]P_i
\]

\[
= \sum_{i=1}^{N} E[d(X, \hat{y}_i)|X \in R_i]P_i
\]

\[
\geq \sum_{i=1}^{N} \min_{y} E[d(X, y)|X \in R_i]P_i
\]

\[
= \sum_{i=1}^{N} E[d(X, y_i)|X \in R_i]P_i = Ed(X, Q(X)) \quad \square
\]
Special Case: \( d(x, y) = (x - y)^2 \).

**Theorem 3 (Centroid Condition for MSE)**

The \( y_i \)'s minimizing the distortion given a partition \( \{R_1, \ldots, R_N\} \) are uniquely given by

\[
y_i = E[X|X \in R_i], \quad i = 1, \ldots, N
\]

**Note:** If \( X \sim f \), then

\[
f_{X|R_i}(x) = \begin{cases} \frac{f(x)}{P(X \in R_i)} & \text{if } x \in R_i \\ 0 & \text{otherwise} \end{cases}
\]

so

\[
E[X|X \in R_i] = \int_{-\infty}^{\infty} x f_{X|R_i}(x) \, dx = \frac{\int_{R_i} x f(x) \, dx}{\int_{R_i} f(x) \, dx}
\]

**Proof of Theorem 3:** Recall that for any r.v. \( Z \) with \( E(Z^2) < \infty \)

\[
E[(Z - E(Z))^2] \leq E[(Z - y)^2]
\]

and equality holds iff \( y = E(Z) \).

Let \( Z_i \) have the conditional distribution of \( X \) given \( X \in R_i \). Then

\[
E(Z_i) = E[X|X \in R_i] = y_i
\]

For any \( y \)

\[
E[(X - y)^2|X \in R_i] = E[(Z_i - y)^2] \\
\geq E[(Z_i - E(Z_i))^2] \quad \text{from (\ast)} \\
= E[(Z_i - y_i)^2] \\
= E[(X - y_i)^2|X \in R_i]
\]

Also, \( y_i \) is the unique minimizer by (\ast). \( \Box \)

**Remarks:**

- The Nearest Neighbor Condition (NNC) and the Centroid Condition (CC) are necessary conditions for optimality.
- An optimal quantizer has to satisfy both the NNC and the CC. The converse is not true: NNC+CC do not imply that a quantizer is optimal.
- NNC+CC = Lloyd-Max conditions. A quantizer satisfying both conditions is called a Lloyd-Max quantizer.
- The NNC determines the optimal encoder \( \mathcal{E} \) given a decoder \( \mathcal{D} \). The CC determines the optimal decoder \( \mathcal{D} \) given an encoder \( \mathcal{E} \).
- The NNC is independent of the source distribution (depends only on the distortion measure).
- The NNC implies that an optimal quantizer can be described by its codebook alone.

**Simple examples for Lloyd-Max quantizers (MSE)**

**Example 1 (2-level quantizer):** Assume \( X \sim f \) and that \( f(x) = f(-x) \) for all \( x \). Let

\[
c = E[X|X > 0] = \frac{\int_{0}^{\infty} x f(x) \, dx}{\int_{0}^{\infty} f(x) \, dx} = 2 \int_{0}^{\infty} x f(x) \, dx
\]

and define

\[
Q(x) = \begin{cases} c & \text{if } x > 0 \\ -c & \text{if } x \leq 0 \end{cases}
\]

Then \( N = 2 \), \( R_1 = (-\infty, 0] \), \( y_1 = -c \), \( R_2 = (0, \infty) \), \( y_2 = c \).

- Check NNC: \( x_1 = \frac{y_1 + y_2}{2} = 0 \) (satisfied)
- Check CC: \( y_i = E[X|X \in R_i], i = 1, 2 \) (satisfied)

**Note:** \( Q \) is a Lloyd-Max quantizer, but not necessarily optimal!
Example cont’d: Special case: assume $X \sim N(0, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}$$

Then

$$c = \frac{2}{\sqrt{2\pi}\sigma} \int_{0}^{\infty} xe^{-x^2/(2\sigma^2)} dx = \frac{2}{\sqrt{2\pi}\sigma} \left[ -\sigma^2 e^{-x^2/(2\sigma^2)} \right]_{0}^{\infty} = \sigma \sqrt{\frac{2}{\pi}}$$

It can be proved that $Q$ is indeed the (unique) optimal 2-level quantizer for the Gaussian case.

Example 2 (Uniform quantizer): $N$-level uniform quantizer over the interval $(a, b)$: for $i = 1, \ldots, N$

$$R_i = (a + (i - 1)\Delta, a + i\Delta)$$

$$y_i = a + \Delta \left( i - \frac{1}{2} \right)$$

- Each $R_i$ is an interval of length $\Delta = \frac{b - a}{N}$
- $y_i$ is the midpoint of $R_i$.
- $x_i = \frac{y_i + y_{i+1}}{2} \implies$ NNC is satisfied.

Assume $X$ is uniformly distributed over $[a, b]$:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

The Lloyd quantizer design algorithm

Idea: For a fixed codebook, optimize the partition. For the resulting partition, optimize the codebook. Iterate...

Lloyd Iteration

(a) Given $C_{m} = \{y_{1}^{(m)}, \ldots, y_{N}^{(m)}\}$, use the NN condition to form the optimal partition

$$R_{i}^{(m)} = \{x : d(x, y_{i}^{(m)}) \leq d(x, y_{j}^{(m)}), \quad j = 1, \ldots, N\} \quad i = 1, \ldots, N$$

(b) Determine $C_{m+1} = \{y_{1}^{(m+1)}, \ldots, y_{N}^{(m+1)}\}$ using the centroid condition

$$y_{i}^{(m+1)} = \arg \min_{y \in \mathbb{R}} E[d(X, y)|X \in R_{i}^{(m)}], \quad i = 1, \ldots, N$$

(For MSE $y_{i}^{(m+1)} = E[X|X \in R_{i}^{(m)}]$)
**Lloyd algorithm**

**Step 1** Inputs: pdf \( f(x) \), initial codebook \( C_1 = \{y_1^{(1)}, \ldots, y_N^{(1)}\} \), threshold \( \epsilon > 0 \). Set \( m = 1 \) and \( D_1 = Ed(X, Q^{(1)}(X)) \).

**Step 2** Given \( C_m \), perform the Lloyd iteration (a) and (b) to generate \( C_{m+1} \).

**Step 3** Compute \( D_{m+1} = Ed(X, Q^{(m+1)}(X)) \).
- If \( \frac{D_m - D_{m+1}}{D_m} < \epsilon \), then output \( C_{m+1} \) and stop.
- Otherwise \( m := m + 1 \) and go to Step 2.

**Remarks:**
- NNC+CC: (a) and (b) in the Lloyd iteration either reduce the distortion or leave it unchanged. Thus
  \[
  D_1 \geq D_2 \geq \cdots \geq D_m \geq D_{m+1} \geq \cdots
  \]
  and \( D_m \) converges to a limit as \( m \to \infty \). Hence
  \[
  \lim_{m \to \infty} \frac{D_m - D_{m+1}}{D_m} = 0
  \]
  and so the algorithm always stops after a finite number of iterations.
- It is not guaranteed that the sequence of codebooks \( C_m \) converges as \( m \to \infty \).
- Even if \( C_m \) converges, the “limit quantizer” may not be optimal (it is only guaranteed to be “locally optimal”).

**Design from training data**

In practice, the distribution (pdf) of \( X \) is unknown. Instead, independent (or dependent) samples \( V_1, V_2, \ldots, V_M \) with \( V_j \sim X \) are given.

(i) We could form an estimate \( f_M(x) \) of \( f(x) \) from \( V_1, \ldots, V_M \). Then use the Lloyd algorithm with \( f_M \). Computationally demanding.

(ii) Direct approach: given the training set \( \{v_1, \ldots, v_M\} \), approximate the distribution of \( X \) by the distribution of a discrete r.v.

\[
P(Z_M = v_j) = \frac{1}{M}
\]

i.e., assign probability mass \( 1/M \) to each \( v_j \).

If the \( v_j \) are all distinct (which happens with prob. 1 if produced by a source with a pdf), then a valid discrete distribution is obtained. This is the empirical distribution of the training sequence.

**Remarks:**
- For any quantizer \( Q \)
  \[
  Ed(Z_M, Q(Z_M)) = \sum_{j=1}^{M} d(v_j, Q(v_j))P(Z_M = v_j)
  = \frac{1}{M} \sum_{j=1}^{M} d(v_j, Q(v_j))
  \]
  Thus optimizing \( Q \) for \( Z_M \) is equivalent to minimizing the average distortion of \( Q \) over the training set.
- Assume \( v_1, \ldots, v_M \) are drawn independently from the distribution of \( X \) or form a stationary and ergodic sequence. It can be shown that as \( M \to \infty \), the distribution of \( Z_M \) (the empirical distribution) approximates the distribution of \( X \) increasingly well. Thus for large \( M \), minimizing the average distortion of \( Q \) over the training set approximates optimizing \( Q \) for \( X \).
**Lloyd-Max conditions for training sets (MSE)**

**NN condition:** Since independent of the source, remains unchanged. Need only to partition the training set \( T = \{ v_1, \ldots, v_M \} \):

\[
R_i = \{ v \in T : |v - y_i| \leq |v - y_j|, j = 1, \ldots, N \} \quad i = 1, \ldots, N
\]

Tie breaking: if \( v \in T \) has two nearest neighbors, assign it to the one with the smaller index.

**Centroid condition:** Given \( R_1, \ldots, R_N \), the MSE centroids are

\[
y_i = E[Z_M | Z_M \in R_i] = \frac{1}{|R_i|} \sum_{v_j \in R_i} v_j
\]

where \(|R_i|\) is the number of training samples in \( R_i \).

---

**Performance Analysis**

Distortion of optimal \( N \)-level quantizer:

\[
D^*(N) = \min_{Q \in \mathcal{Q}_N} Ed(X, Q(X))
\]

For \( X \sim f \) and MSE

\[
D^*(N) = \min_{Q \in \mathcal{Q}_N} \int_{-\infty}^{\infty} (x - Q(x))^2 f(x) \, dx
\]

\[
= \min_{y_1 < y_2 < \cdots < y_N} \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} (x - y_i)^2 f(x) \, dx
\]

where \( x_i = \frac{1}{2}(y_i + y_{i+1}), \quad i = 1, \ldots, N - 1 \).

Thus determining \( D^*(N) \) and the optimal \( Q^* \) involves the minimization of a real function of \( N \) real variables. Very hard, nonlinear and non-convex problem!

---

**Lloyd algorithm for training data and MSE**

**Step 1** Inputs: training set \( T = \{ v_1, \ldots, v_M \} \), initial codebook \( C_1 = \{ y_1^{(1)}, \ldots, y_N^{(1)} \} \), and threshold \( \epsilon > 0 \). Set \( m = 1 \).

**Step 2** Given \( C_m \), partition \( T \) into \( N \) bins \( R_1^{(m)}, \ldots, R_N^{(m)} \) using the NNC. Calculate the empirical average inside each bin

\[
y_i^{(m+1)} = \frac{1}{|R_i^{(m)}|} \sum_{v_j \in R_i^{(m)}} v_j
\]

to generate \( C_{m+1} \).

**Step 3** Compute

\[
D_{m+1} = \frac{1}{M} \sum_{j=1}^{M} (v_j - Q(v_j))^2
\]

If \( (D_m - D_{m+1})/D_m < \epsilon \), then output \( C_{m+1} \) and stop. Otherwise \( m := m + 1 \) and go to **Step 2**.

---

**Companding quantization**

Provides structure facilitating performance analysis

\[
X \xrightarrow{G} Q_\Delta \xrightarrow{G^{-1}} Q(x)
\]

- \( G : \mathbb{R} \to (a, b) \) monotone increasing and invertible (compressor)
- \( Q_\Delta : N \)-level uniform quantizer with support \((a, b)\) and step size \( \Delta = \frac{b-a}{N} \).
- \( G^{-1} : (a, b) \to \mathbb{R} \), the inverse of \( G \) (expander)

\[
Q(x) = G^{-1}(Q_\Delta(G(x)))
\]
**Proposition 1**

For any $N$-level regular quantizer $Q$ there exists a companding realization.

---

**Proof of Proposition 1:** For simplicity assume $Q$ has bounded support. The output levels and decision levels are ordered as $x_0 < y_1 < x_1 < y_2 < \cdots < y_N < x_N$.

Prescribe the values of $G$ such that

- $G(y_i) = a + \Delta(i - \frac{1}{2})$, $i = 1, \ldots, N$
- $G(x_i) = a + i \Delta$, $i = 0, \ldots, N$
- $G(x)$ is *linearly* interpolated between these discrete points

Then

- $G(x) : (x_0, x_N) \rightarrow (a, b)$ is strictly monotone increasing (so it is invertible)
- $x \in (x_{i-1}, x_i) \Rightarrow G(x) \in (a + (i - 1)\Delta, a + i\Delta) \Rightarrow Q_\Delta(G(x)) = a + \Delta(i - \frac{1}{2}) \Rightarrow G^{-1}(Q_\Delta(G(x))) = y_i$ \hfill $\Box$

**Note:** $G(x)$ can be “smoothed out” while retaining these properties.

---

**Example:** $N = 5$

---

**Another look at $D^*(N)$**

- Fix $(a, b)$ and let $Q_{G,N}$ denote the $N$-level quantizer realized by the compressor function $G$.
- By the proposition
  \[ D^*(N) = \min_{Q \in \mathcal{Q}_N} E[(X - Q(X))^2] = \min_G E[(X - Q_{G,N}(X))^2] \]
- Hence it is of great interest to determine $\min_G E[(X - Q_{G,N}(X))^2]$
- Will do this under the assumption that $N$ is large (“high-resolution conditions”)
High-resolution performance

Fix \( G \), let \( N \rightarrow \infty \)

Assumptions

(i) \( N \) is large, \( \Delta_i = (x_i - x_{i-1}) \) is small for \( i = 2, \ldots, N - 1 \)
(ii) \( X \sim f \), where \( f \) is a continuous pdf
(iii) \( G : \mathbb{R} \rightarrow (a, b) \) is continuously differentiable

Will do a series of approximations based on these assumptions

\[
D(Q_{G,N}) = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} (x - y_i)^2 f(x) \, dx
\]

\[
= \sum_{i=2}^{N-1} \int_{x_{i-1}}^{x_i} (x - y_i)^2 f(x) \, dx + \int_{x_{N-1}}^{x_N} (x - y_N)^2 f(x) \, dx
\]

“granular” distortion \( D_{gr} \)

\[
+ \int_{-\infty}^{x_1} (x - y_1)^2 f(x) \, dx + \int_{x_N}^{\infty} (x - y_N)^2 f(x) \, dx
\]

“overload” distortion \( D_{ol} \)

Assumption (iv): The overload distortion is negligible compared to the granular distortion

\( D(Q) \approx D_{gr} \)

Thus

\[
\int_{x_{i-1}}^{x_i} (x - y_i)^2 f(x) \, dx \approx f(y_i) \int_{x_{i-1}}^{x_i} (x - y_i)^2 \, dx
\]

(since \( f \) is continuous and \( x_i - x_{i-1} \) is small)

\[
\approx f(y_i) \int_{x_{i-1}}^{x_i} \left( x - \frac{x_i - x_{i-1} + x_i}{2} \right)^2 \, dx
\]

\[
= f(y_i) \frac{(x_i - x_{i-1})^3}{12}
\]

\[
\approx \frac{1}{12} f(y_i) \Delta^2 G'(y_i)^2 (x_i - x_{i-1})
\]

\[
= \frac{1}{12} f(y_i) \frac{B^2}{N^2} G'(y_i)^2 (x_i - x_{i-1})
\]

(by letting \( B = (b - a) \) so that \( \Delta = \frac{B}{N} \))
Putting things together

\[ D(Q_{G,N}) \approx \sum_{i=2}^{N-1} \int_{x_{i-1}}^{x_i} (x - y_i)^2 f(x) \, dx \]

\[ \approx \frac{B^2}{12N^2} \sum_{i=2}^{N-1} \frac{f(y_i)}{G'(y_i)^2} (x_i - x_{i-1}) \]

(Riemann sum)

\[ \approx \frac{B^2}{12N^2} \int_{x_1}^{x_{N-1}} f(x) \, dx \]

\[ \approx \frac{B^2}{12N^2} \int_{-\infty}^{\infty} f(x) \, dx \]

**Conclusion:** For MSE and \( N \) large

\[ D(Q_{G,N}) \approx \frac{B^2}{12N^2} \int_{-\infty}^{\infty} f(x) \, dx \]

(Here \( \approx \) means that the ratio of the two sides converges to 1 as \( N \to \infty \))

Easy-to-calculate, analytic formula!

**Example:** Let \( X \sim f \) such that \( f(x) = 0 \) if \( x \notin (a, b) \). Consider the \( N \)-level uniform quantizer \( Q_\Delta \) over \((a, b)\), where \( \Delta = (b - a)/N \).

If \( G : (a, b) \to (a, b) \) is defined as \( G(x) = x \), then for all \( x \in (a, b) \)

\[ Q_{N,G}(x) = G^{-1}(Q_\Delta(G(x))) = Q_\Delta(x) \]

From the compander approximation, for large \( N \)

\[ D(Q_\Delta) = D(Q_{G,N}) \approx \frac{(b - a)^2}{12N^2} \int_{-\infty}^{\infty} \frac{f(x)}{G'(x)^2} \, dx \]

\[ = \frac{1}{12} \left( \frac{b - a}{N} \right)^2 \int_{a}^{b} f(x) \, dx \] (since \( G'(x) = 1 \))

\[ = \frac{\Delta^2}{12} \]

**Intuitive explanation:** If \( N \) is large, \( \Delta \) is small, and the pdf is approximately uniform over each quantizer cell

**Optimal asymptotic performance**

- We want to find the compressor \( G \) minimizing \( D(Q_{G,N}) \) for large \( N \).
- Any \( N \)-level quantizer can be represented as a companding quantizer with compressor function \( G : \mathbb{R} \to (0,1) \).
  \[ \Rightarrow \text{It is enough to consider compressors } G : \mathbb{R} \to (0,1) \]
- In this case (since \( B = 1 \))

\[ D(Q_{G,N}) \approx \frac{1}{12N^2} \int_{-\infty}^{\infty} \frac{f(x)}{G'(x)^2} \, dx \]

- We want to minimize

\[ \int_{-\infty}^{\infty} \frac{f(x)}{G'(x)^2} \, dx \]
Theorem 4 (Optimal compressor)

Given the pdf \( f \), the compressor \( G : \mathbb{R} \to (0, 1) \) which minimizes
\[
\int_{-\infty}^{\infty} \frac{f(x)}{G'(x)^2} \, dx
\]
is determined by
\[
G'(x) = \frac{f(x)^{1/3}}{\int_{-\infty}^{\infty} f(y)^{1/3} \, dy}
\]
For the (asymptotically) optimal companding scheme using this \( G \)
\[
D(Q_{G,N}) \approx \frac{1}{12N^2} \left( \int_{-\infty}^{\infty} f(x)^{1/3} \, dx \right)^3
\]

Proof of Lemma 5: First we note that the strict concavity of \( \log(x) \)
implies that for any \( a, b \geq 0 \) and \( 0 < \alpha < 1 \),
\[
a^\alpha b^{1-\alpha} \leq \alpha a + (1-\alpha)b
\]
where equality holds iff \( a = b \). (Proof: if \( ab > 0 \), take the logarithm of both sides.)
Define
\[
a(x) = \left( \frac{u(x)}{\|u\|_p} \right)^p, \quad b(x) = \left( \frac{v(x)}{\|v\|_q} \right)^q
\]
Let \( \alpha = \frac{1}{p} \) (so that \( 1 - \alpha = \frac{1}{q} \)). From (*) we have
\[
\frac{u(x) \cdot v(x)}{\|u\|_p \cdot \|v\|_q} = a(x)^\alpha b(x)^{1-\alpha}
\]
\[
\leq \alpha a(x) + (1-\alpha)b(x)
\]
\[
= \alpha \left( \frac{u(x)}{\|u\|_p} \right)^p + (1-\alpha) \left( \frac{v(x)}{\|v\|_q} \right)^q
\]
Now integrate both sides.

Lemma 5 (Hölder’s inequality)

Assume \( p > 1 \) and \( q > 1 \) are such that \( \frac{1}{p} + \frac{1}{q} = 1 \). Let \( u(x) \geq 0 \) and \( v(x) \geq 0 \) satisfy
\[
\int_{-\infty}^{\infty} u(x)^p \, dx < \infty \quad \int_{-\infty}^{\infty} v(x)^q \, dx < \infty
\]
Then
\[
\int_{-\infty}^{\infty} u(x)v(x) \, dx \leq \left( \int_{-\infty}^{\infty} u(x)^p \, dx \right)^{1/p} \left( \int_{-\infty}^{\infty} v(x)^q \, dx \right)^{1/q}
\]
Moreover, equality holds iff \( v(x)^q = Cu(x)^p \) for some \( C > 0 \).

Remark: For any \( p > 0 \) and function \( g : \mathbb{R} \to \mathbb{R} \) define
\[
\|g\|_p = \left( \int_{-\infty}^{\infty} |g(x)|^p \, dx \right)^{1/p}
\]
Then Hölder’s inequality states that
\[
\int_{-\infty}^{\infty} u(x)v(x) \, dx \leq \|u\|_p \|v\|_q
\]

Proof of Lemma cont’d:
\[
\int_{-\infty}^{\infty} \frac{u(x) \cdot v(x)}{\|u\|_p \cdot \|v\|_q} \, dx
\]
\[
\leq \alpha \int_{-\infty}^{\infty} \left( \frac{u(x)}{\|u\|_p} \right)^p \, dx + (1-\alpha) \int_{-\infty}^{\infty} \left( \frac{v(x)}{\|v\|_q} \right)^q \, dx
\]
\[
= \alpha + (1-\alpha) = 1
\]
Hence
\[
\int_{-\infty}^{\infty} u(x)v(x) \, dx \leq \|u\|_p \|v\|_q
\]
Note that equality holds if and only if \( a(x) = b(x) \) for (almost) all \( x \), i.e.,
iff
\[
\left( \frac{u(x)}{\|u\|_p} \right)^p = \left( \frac{v(x)}{\|v\|_q} \right)^q
\]
This is equivalent to \( v(x)^q = Cu(x)^p \) for some \( C > 0 \). \( \square \)
Proof of Thm. 4: Let $G : \mathbb{R} \to (0, 1)$ be an arbitrary compressor. Define

$$u(x) = \left( \frac{f(x)}{G(x)^2} \right)^{1/3} \quad v(x) = G'(x)^{2/3}$$

and let $p = 3$, $q = 3/2$ (so that $\frac{1}{p} + \frac{1}{q} = 1$).

By Hölder’s inequality

$$\int_{-\infty}^{\infty} \left( \frac{f(x)}{G'(x)^2} \right)^{1/3} G'(x)^{2/3} \, dx \leq \left( \int_{-\infty}^{\infty} \frac{f(x)}{G'(x)^2} \, dx \right)^{1/3} \left( \int_{-\infty}^{\infty} G'(x) \, dx \right)^{2/3} = 1 \quad (\ast)$$

(\ast) follows since $G$ is monotone increasing and maps onto $(0, 1)$ so

$$\int_{-\infty}^{\infty} G'(x) \, dx = \lim_{y \to \infty} \left[ G(x) \right]_y^\infty = \lim_{y \to \infty} G(y) - \lim_{y \to -\infty} G(y) = 1 - 0 = 1$$

Proof of Thm. 4 cont’d:

We showed

$$\int_{-\infty}^{\infty} \frac{f(x)}{G'(x)^2} \, dx \geq \left( \int_{-\infty}^{\infty} f(x)^{1/3} \, dx \right)^3 \quad (\ast\ast)$$

From Lemma 5, the lower bound (\ast\ast) is achieved by $G$ iff

$$G'(x) = C \frac{f(x)}{G'(x)^2} \iff G'(x) = \hat{C} f(x)^{1/3}$$

Since $\int_{-\infty}^{\infty} G'(x) \, dx = 1$, we get $\hat{C} = \left( \int_{-\infty}^{\infty} f(x)^{1/3} \, dx \right)^{-1}$

Thus the optimal compressor satisfies

$$G'(x) = \frac{f(x)^{1/3}}{\int_{-\infty}^{\infty} f(y)^{1/3} \, dy}$$

\[ \square \]

The 6 dB per bit rule

- Consider the SNR as a function of $R = \log_2 N$. Since $N^{-2} = 2^{-2R}$

$$D^*(N) \approx \frac{1}{N^2} \frac{1}{12} \| f \|_{1/3} = 2^{-2R} C(f)$$

where $C(f) = \frac{1}{12} \| f \|_{1/3}$ does not depend on $N$.

- Thus the SNR in dB is

$$\text{SNR} \approx 10 \log_{10} \left( \frac{E(X^2)}{2^{-2R} C(f)} \right) = R \cdot 20 \log_{10} 2 + 10 \log_{10} \left( \frac{E(X^2)}{C(f)} \right) \approx 6.02$$

- **Rule of thumb**: increasing the rate by 1 bit results in a 6 dB increase in SNR.

We have “proved” that for continuous sources with a pdf and $N$ large

$$D^*(N) = \min_G D(Q_{G,N}) \approx \frac{1}{12 N^2} \| f \|_{1/3}$$

The exact result (which is much harder to prove) is

**Theorem 6 (Bucklew and Wise)**

Assume the source pdf satisfies $\int_{-\infty}^{\infty} |x|^{2+\epsilon} f(x) \, dx < \infty$ for some $\epsilon > 0$. Then

$$\lim_{N \to \infty} N^2 D^*(N) = \frac{1}{12} \| f \|_{1/3}$$