Data Compression and Source Coding
VI: Vector Quantization

T. Linder
Queen’s University
Winter 2016

**Definition** A $k$-dimensional $N$-point vector quantizer (VQ) is a mapping

$$Q : \mathbb{R}^k \to C$$

where $C = \{c_1, c_2, \ldots, c_N\} \subset \mathbb{R}^k$.

- $C$ is called the **codebook**.
- $c_1, \ldots, c_N$ are the **output levels** (or output points, reproduction points, code points, etc).
- $R_i = \{x : Q(x) = c_i\}$, $i = 1, \ldots, N$, are the quantizer cells.

As in the scalar case, the cells form a partition of $\mathbb{R}^k$:

$$R_i \cap R_j = \emptyset \quad \text{if} \quad i \neq j \quad \text{and} \quad \bigcup_{i=1}^{N} R_i = \mathbb{R}^k$$

**Example 1**: Product VQ.

- $Q_1$: $N_1$-level scalar quantizer with codebook $C_1$ and cells $\{R_i^{(1)}\}$
- $Q_2$: $N_2$-level scalar quantizer with codebook $C_2$ and cells $\{R_j^{(2)}\}$
- For any $x = (x_1, x_2)$, define $Q(x)$ by
  $$Q(x) = (Q_1(x_1), Q_2(x_2))$$

Then $Q$ is a special 2-dimensional vector quantizer with codebook

$$C = C_1 \times C_2 = \{c = (c_1, c_2) : c_1 \in C_1, c_2 \in C_2\}$$

and cells

$$R_{ij} = R_i^{(1)} \times R_j^{(2)}, \quad i = 1, \ldots, N_1, \quad j = 1, \ldots, N_2$$

- Note that $Q$ has $|C| = N_1 N_2$ output points.

**Example 2**: Transform coder.

- $x = (x_1, \ldots, x_k)^t$, $Ax = y = (y_1, \ldots, y_k)^t$
- $\hat{y} = (Q_1(y_1), \ldots, Q_k(y_k))^t$, $A^{-1}\hat{y} = \hat{x} = (\hat{x}_1, \ldots, \hat{x}_k)^t$

Each $Q_i(y_i)$ can take $N_i$ distinct values $\Rightarrow \hat{y}$ can take $\prod_{i=1}^{k} N_i$ distinct values $\Rightarrow \hat{x}$ can take $\prod_{i=1}^{k} N_i$ distinct values (since $A^{-1}$ is nonsingular).

Thus the mapping $Q(x) = \hat{x}$ is a $k$-dimensional vector quantizer with $\prod_{i=1}^{k} N_i$ output points.
Encoder-Decoder structure

- Encoder: $E: \mathbb{R}^k \to \{1, \ldots, N\} = \mathcal{I}$ defined by $E(x) = j \iff x \in R_j$
- Decoder: $D: \{1, \ldots, N\} \to \{c_1, \ldots, c_N\}$ defined by $D(j) = c_j$

Thus $Q(x) = D(E(x))$

$x \xrightarrow{E} j \in \mathcal{I} \xrightarrow{D} Q(x)$

- Rate of $Q$:
  $$r_F(Q) = \frac{1}{k} \log_2 N \text{ bits/sample}$$

  (The subscript $F$ indicates that the indices $j \in \mathcal{I}$ are encoded using binary strings of fixed-length.)

- Note: Unlike in scalar quantization, the rate can be less than 1.

Some distortion measures of interest:

- MSE: for $x = (x_1, \ldots, x_k)^t$ and $y = (y_1, \ldots, y_k)^t$
  $$d(x, y) = \sum_{i=1}^k (x_i - y_i)^2 = \|x - y\|^2$$

- $r$th power distortion:
  $$d(x, y) = \sum_{i=1}^k |x_i - y_i|^r, \quad r > 0$$

  (absolute error for $r = 1$, MSE for $r = 2$).

- Weighted squared error:
  $$d(x, y) = (x - y)^t W(x - y)$$

  where $W$ is a symmetric and positive definite $k \times k$ matrix.

VQ distortion

- Distortion measure: $d(x, y) \geq 0$ for all $x, y \in \mathbb{R}^k$.
- For a vector random variable $X = (X_1, \ldots, X_k)^t$, the expected distortion of $Q$ is
  $$D(Q) = \mathbb{E}[d(X, Q(X))]$$
  $$= \int_{\mathbb{R}^k} d(x, Q(x)) f(x) \, dx \quad \text{(if } X \sim f)$$
  $$= \sum_{j=1}^N \int_{R_j} d(x, c_j) f(x) \, dx$$

Polytopal VQ

Definitions: A half-space in $\mathbb{R}^k$ is a set of the form

$$H(u, \beta) = \{x \in \mathbb{R}^k: u^t x + \beta \geq 0\}$$

where $u \in \mathbb{R}^k$ ($u \neq 0$) and $\beta \in \mathbb{R}$.

A hyperplane $\Lambda(u, \beta)$ is the boundary of $H(u, \beta)$:

$$\Lambda(u, \beta) = \{x: u^t x + \beta = 0\}$$

A set $R \subset \mathbb{R}^k$ is a convex polytope if it can be written as a finite intersection of half-spaces, i.e.,

$$R = \bigcap_{i=1}^L H(u_i, \beta_i)$$

for some $u_1, \ldots, u_L \in \mathbb{R}^k$ and $\beta_1, \ldots, \beta_L \in \mathbb{R}$. 
Reminder: A set \( S \subseteq \mathbb{R}^k \) is convex if for all \( x, y \in S \) and \( 0 < \alpha < 1 \),
\[
\alpha x + (1 - \alpha)y \in S
\]

**Claim:** A convex polytope is a convex subset of \( \mathbb{R}^k \).

**Proof:** Let \( R = \bigcap_{i=1}^L H(u_i, \beta_i) \). Pick any \( x, y \in H(u_i, \beta_i) \) and \( \alpha \in (0, 1) \). Then
\[
u_i^t(\alpha x + (1 - \alpha)y) + \beta_i = \alpha(\nu_i^tx + \beta_i) + (1 - \alpha)(\nu_i^ty + \beta_i) \geq 0
\]
Thus \( \alpha x + (1 - \alpha)y \in H(u_i, \beta_i) \), so each \( H(u_i, \beta_i) \) is convex.

Now if \( x, y \in R \), then \( x, y \in H(u_i, \beta_i) \) for all \( i \). Let \( \alpha \in (0, 1) \). Then
\[
\alpha x + (1 - \alpha)y \in H(u_i, \beta_i)
\]
so that \( \alpha x + (1 - \alpha)y \in \bigcap_{i=1}^L H(u_i, \beta_i) = R \). Hence \( R \) is convex. \( \square \)

**Definition:** \( Q \) is a nearest neighbor (NN) vector quantizer if for all \( x \)
\[
Q(x) = \arg \min_{c_j \in C} d(x, c_j)
\]
or equivalently, for all \( i = 1, \ldots, N \)
\[
R_i \subseteq \{ x : d(x, c_i) \leq d(x, c_j), \ j = 1, \ldots, N \}
\]

**Claim:** For the squared error distortion measure, the overlapping quantization regions
\[
\tilde{R}_i = \{ x : d(x, c_i) \leq d(x, c_j), \ j = 1, \ldots, N \}
\]
of a nearest neighbor quantizer are convex polytopes.

**Remark:** The polytopes \( \tilde{R}_1, \ldots, \tilde{R}_N \) are called the Voronoi regions corresponding to the points \( c_1, \ldots, c_N \).

**Proof of claim:** For \( d(x, y) = \|x - y\|^2 \),
\[
\tilde{R}_i = \{ x : d(x, c_i) \leq d(x, c_j), \ j = 1, \ldots, N \}
\]
\[
= \bigcap_{j=1}^N \{ x : \|x - c_i\|^2 \leq \|x - c_j\|^2 \}
= \bigcap_{j=1}^N H_{ij}
\]
Note that
\[
\|x - c_j\|^2 = (x - c_j)^t(x - c_j)
= \|x\|^2 + \|c_j\|^2 - 2c_j^tx
\]
Hence
\[
H_{ij} = \{ x : \|x - c_i\|^2 \leq \|x - c_j\|^2 \}
= \{ x : \|x\|^2 + \|c_i\|^2 - 2c_i^tx \leq \|x\|^2 + \|c_j\|^2 - 2c_j^tx \}
= \{ x : -2(c_i - c_j)^tx + \|c_i\|^2 - \|c_j\|^2 \leq 0 \}
\]
\[
= \{ x : -u_{ij}^tx + \beta_{i,j} \geq 0 \}
\]
Thus each \( H_{ij} \) is a half-space. Since \( \tilde{R}_i \) is the intersection of the \( H_{ij} \),
\( j = 1, \ldots, N \), it is a convex polytope. \( \square \)
Remarks:
- A \((k-1)\)-dimensional face of the polytope \(\tilde{R}_i\) is the part of the hyperplane
  \[ \{ x : \| x - c_i \|^2 = \| x - c_j \|^2 \} = \{ x : u_{ij}^T x + \beta_{ij} = 0 \} \]
  that belongs to \(\tilde{R}_i\). (Thus \(\tilde{R}_i\) has at most \(N-1\) faces.)
- If \(c_i\) and \(c_j\) are both nearest neighbors of \(x\), i.e,
  \[ \| x - c_i \| = \| x - c_j \| = \min_{1 \leq k \leq N} \| x - c_k \| \]
  then a tie-breaking rule is needed to decide if \(x \in R_i\).
- In geometric terms, this means that depending on the tie-breaking rule, any of the \((k-1)\)-dimensional faces of the polytope \(\tilde{R}_i\) may or may not belong to \(R_i\).

VQ optimality conditions
The notion of optimality is the same as in the scalar case:

Definition: Let \(Q_N\) denote the family of all \(k\)-dimensional \(N\)-level quantizers. \(Q^* \in Q_N\) is an optimal quantizer for source \(X\) if
\[ E[d(X, Q^*(X))] = \min_{Q \in Q_N} E[d(X, Q(X))] \]

Remarks:
- Just as in the scalar case, \(Q^*\) depends on the distribution of \(X\) and on \(d\).
- Also, \(Q^*\) is not necessarily unique, but it always exists for any well-behaved distortion measure \(d\) if \(E[d(X, y)] < \infty\) for some \(y \in \mathbb{R}^k\).

Definition (Regular VQ): \(Q\) is called regular if
- (a) each \(R_i\) is a convex subset of \(\mathbb{R}^k\),
- (b) \(c_i \in R_i\) for all \(i = 1, \ldots, N\).

Note: For the squared error distortion measure, any nearest neighbor quantizer has convex cells. Also, \(\| c_i - c_i \| < \| c_i - c_j \|\) for all \(j \neq i\), so we always have \(c_i \in R_i\).
\(\Rightarrow\) Nearest neighbor quantizers are regular.

Theorem 1 (Nearest Neighbor Condition)
Any nearest neighbor quantizer has minimum distortion among all \(N\)-level vector quantizers with the same codebook.

The proof of the statement is the same as in the scalar case.
**Theorem 2 (Centroid Condition)**

Consider all \( N \)-level vector quantizers with given cells \( R_1, \ldots, R_N \). Among these, the quantizer \( Q \) with output points

\[
c_i = \arg \min_{c \in \mathbb{R}^k} E[d(X, c) | X \in R_i], \quad i = 1, \ldots, N
\]

has minimum distortion.

Writing out the expected distortion as

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

the same proof as in the scalar case applies. The centroid \( c_i \) may not be unique.

**Special Case:** \( d(x, y) = ||x - y||^2 \).

**Theorem 3**

For the squared error distortion measure, the centroid of \( R_i \) is uniquely given by

\[
c_i = E[X | X \in R_i]
\]

The theorem follows from the fact that if the random vector \( Z \) has finite second moment and expectation \( c^* = E[Z] \), then for all \( c \in \mathbb{R}^k \)

\[
E[||Z - c||^2] \leq E[||Z - c^*||^2]
\]

where equality holds if and only if \( c = c^* \).

**Note:** If \( X \) has pdf \( f \), then the MSE centroid of cell \( R_i \) is

\[
c_i = \int_{R_i} x f_{X|R_i}(x) \, dx = \frac{\int_{R_i} x f(x) \, dx}{\int_{R_i} f(x) \, dx}
\]

since

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

- Just as in the scalar case, the CC and the NNC are necessary, but not sufficient conditions for the the optimality of an \( N \)-level vector quantizer.
- Any vector quantizer satisfying both the CC and the NNC is called a Lloyd-Max vector quantizer.

**Claim:** For MSE, a Lloyd-Max vector quantizer is regular.

**Proof:** Lloyd-Max quantizers are nearest neighbor (NN) quantizers. We have seen that for the MSE, any NN vector quantizer is a regular quantizer.
With training data \( T = \{ v_1, v_2, \ldots, v_M \} \) instead of the pdf of \( X \), and with MSE distortion, the NN condition is

\[ R_i = \{ v \in T : \| v - c_i \| \leq \| v - c_j \|, j = 1, \ldots, N \} \]

(tie-breaking is needed).

The centroid condition becomes

\[ c_i = \frac{1}{|R_i|} \sum_{v_j \in R_i} v_j \]

The Lloyd VQ design algorithm

The algorithm is called the generalized Lloyd algorithm (GLA) or the Linde-Buzo-Gray (LBG) algorithm.

As in the scalar case, the algorithm stops in a finite number of steps. The resulting VQ is not guaranteed to be optimal (Lloyd-Max quantizers are only “locally optimal”).

In practice, almost exclusively the training set version of the Lloyd algorithm is used (it is difficult to evaluate multiple integrals over polytopes in higher dimensions).

The training set version is closely related to k-means clustering in statistics.

The Lloyd VQ design algorithm

Idea: same as in the scalar case

Lloyd Iteration

(a) Given \( C_m = \{ c_1^{(m)}, \ldots, c_N^{(m)} \} \), use the NN condition to form the optimal partition cells \( R_1^{(m)}, \ldots, R_N^{(m)} \).

(b) Determine \( C_{m+1} = \{ c_1^{(m+1)}, \ldots, c_N^{(m+1)} \} \) as the centroids of \( R_1^{(m)}, \ldots, R_N^{(m)} \).

Note: The expected distortion decreases (or remains unchanged) in each iteration.

The Lloyd algorithm for VQ design is essentially the same as in the scalar case:

Step 1 Inputs: pdf or training set, initial codebook \( C_1 = \{ c_1^{(1)}, \ldots, c_N^{(1)} \} \), threshold \( \epsilon > 0 \). Set \( m = 1 \).

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

Step 3 Compute distortion \( D_{m+1} \) for codebook \( C_{m+1} \).

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.

The Lloyd algorithm is essentially the same as in the scalar case.
How is the initial codebook chosen? Some possibilities are:

- Random choice: generate $N$ vectors from some distribution.
- Use the first $N$ vectors from the training set.
- Pick $N$ vectors at random from the training set.
- Pick $N$ vectors from a $k$-dimensional lattice (will study lattices later).

A useful variant of the Lloyd algorithm is the *splitting algorithm*:

- Design an optimal 0-rate (1-level) VQ by finding the mean of the training set.
- Split the code vector into 2 vectors (there are several ways of doing this), and run the Lloyd algorithm until convergence. This gives a rate-1 code.
- Split each code vector of the rate-1 code to get 4 new vectors. Run the Lloyd algorithm to obtain an optimized rate-2 code.
- Continue splitting and running the Lloyd algorithm in this fashion until the desired rate is reached.

**Lattice vector quantization**

Let $\{u_1, u_2, \ldots, u_k\}$ be linearly independent vectors in $\mathbb{R}^k$.

**Definition** A $k$-dimensional lattice with basis $\{u_1, u_2, \ldots, u_k\}$ is the infinite discrete set

$$\Lambda = \{y \in \mathbb{R}^k : y = \sum_{i=1}^{k} n_i u_i, \ n_1, \ldots, n_k \text{ are integers}\}$$

($\Lambda$ is the collection of integer-coefficient linear combinations of the basis vectors.)

**Note:** It is easy to see that $\Lambda$ is an additive group for the usual addition in $\mathbb{R}^k$:

- $0 \in \Lambda$ ($\Lambda$ contains the origin). Also, $y \in \Lambda \Rightarrow -y \in \Lambda$.
- $y_1, y_2 \in \Lambda \Rightarrow y_1 + y_2 \in \Lambda$.

**Examples:**

- $k = 1$: *Integer lattice* $\Lambda = \mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$.
  
  **Note:** Any other lattice in $\mathbb{R}$ is a scaled version of $\mathbb{Z}$.

- $k = 2$:
  - 2D integer lattice $\mathbb{Z}^2$: $u_1 = (1, 0)^t$, $u_2 = (0, 1)^t$.
  - Hexagonal lattice: $u_1 = (1, 0)^t$, $u_2 = \left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right)^t$.

**Definition:** The set

$$R_0 = \{x \in \mathbb{R}^k : ||x|| \leq ||x - y||, \ y \in \Lambda\}$$

is the basic cell of the lattice $\Lambda$.

**Note:** It can be shown that $R_0$ is a convex polytope.
**Definition:** A lattice vector quantizer is a nearest neighbor quantizer $Q_A$ with codebook $\Lambda$.

- Lattice VQ encoding rule:
  \[ Q_A(x) = \arg \min_{y \in \Lambda} ||x - y||. \]
  (Tie-breaking is needed.)
- The quantizer cell associated with a $y_i \in \Lambda$ is
  \[ R_i = \{ x \in \mathbb{R}^k : ||x - y_i|| \leq ||x - y_j||, \quad y_j \in \Lambda \}. \]
  (The $R_i$ overlap at their boundary.)

**Note:** In general $Q_A$ has an infinite number of code points. If the source $X$ has a bounded support, then only a finite number of these are used, so $Q_A$ has finite rate. Otherwise, we’ll use variable-rate coding for the output of $Q_A$. (Will return to this problem later.)

**Distortion of Lattice VQ**

Assume MSE and $X \sim f$. Then
\[ D(Q_A) = \sum_i \int_{R_i} ||x - y_i||^2 f(x) \, dx. \]

**Note:** Since each $R_i$ is congruent to $R_0$,
\[ \max_i \max_{x \in R_i} ||x - y_i||^2 = \max_{x \in R_0} ||x||^2 \leq \text{diam}(R_0)^2 \]
where $\text{diam}(R_0) = \max_{x, y \in R_0} ||x - y||$ is the diameter of $R_0$.

Thus
\[ D(Q_A) \leq \text{diam}(R_0)^2 \sum_i \int_{R_i} f(x) \, dx = \text{diam}(R_0)^2 \]
so $D(Q_A)$ is always finite.

**Claim:** The cell $R_i$ is the basic cell $R_0$ shifted by $y_i$:
\[ R_i = R_0 + y_i = \{ x + y_i : x \in R_0 \}. \]

**Proof:** Since
\[ R_i = \{ x \in \mathbb{R}^k : ||x - y_i|| \leq ||x - y_j||, \quad y_j \in \Lambda \} \]
we have
\[ x \in R_i \iff ||x - y_i|| \leq ||x - y_j|| \quad \text{for all } y_j \in \Lambda \]
\[ \iff ||x - y_i|| \leq ||(x - y_i) - (y_j - y_i)|| \quad \text{for all } y_j \in \Lambda \]
Note that $\{y_j - y_i : y_j \in \Lambda\} = \Lambda$. Thus
\[ x \in R_i \iff ||x - y_i|| \leq ||(x - y_i) - y_k|| \quad \text{for all } y_k \in \Lambda \]
\[ \iff x - y_i \in R_0 \]
\[ \iff x \in R_0 + y_i \]
\[ \square \]

Assume
- $f(x)$ is smooth.
- $R_0$ is “small” (i.e. $\text{diam}(R_0)$ is small).

Then $|f(x) - f(y)| \approx 0$ for all $x, y \in R_i$, so
\[ \int_{R_i} f(x) \, dx \approx V(R_i) f(y_i) \quad \Rightarrow \quad f(y_i) \approx \frac{1}{V(R_i)} \int_{R_i} f(x) \, dx \]
where $V(R_i)$ is the $k$-dimensional volume of $R_i$. (Note that $V(R_i) = V(R_0)$ for all $i$.)

Also, since $R_i = R_0 + y_i$,\[ \int_{R_i} ||x - y_i||^2 \, dx = \int_{R_0} ||x||^2 \, dx \]
Thus

\[D(Q_\Lambda) = \sum_i \int_{R_i} \|x - y_i\|^2 f(x) \, dx\]

\[\approx \sum_i \int_{R_i} \|x - y_i\|^2 \, dx\]

\[\approx \sum_i \frac{P(X \in R_i)}{V(R_i)} \int_{R_i} \|x - y_i\|^2 \, dx\]

\[= \frac{1}{V(R_0)} \sum_i \frac{P(X \in R_i)}{V(R_i)} \int_{R_0} \|x\|^2 \, dx\]

\[= \frac{1}{V(R_0)} \int_{R_0} \|x\|^2 \, dx\]

**Conclusion:** For smooth \(f(x)\) and if the basic cell is small enough,

\[D(Q_\Lambda) \approx \frac{1}{V(R_0)} \int_{R_0} \|x\|^2 \, dx\]

**Note:**

\[H(R_0) \triangleq \frac{1}{V(R_0)} \int_{R_0} \|x\|^2 \, dx\]

is the moment of inertia of the \(k\)-dimensional convex polytope \(R_0\) (This physical interpretation assumes that \(R_0\) has unit mass that is uniformly distributed on it).

**Example:** The uniform quantizer \(Q_\Delta\) with step size \(\Delta\) and quantization levels \(0, \pm \Delta, \pm 2\Delta, \ldots\) is a lattice quantizer with basic cell \(R_0 = [-\Delta/2, \Delta/2]\) and distortion

\[D(Q_\Delta) \approx \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} x^2 \, dx = \frac{\Delta^2}{12}\]

as we have seen before.

- Since \(D(Q_\Lambda) \approx H(R_0)\), the smaller \(H(R_0)\), the lower the distortion of \(Q_\Lambda\). But \(H(R_0)\) decreases if \(\Lambda\) is simply rescaled by a factor \(0 < \alpha < 1\).

- Need a scale-invariant quantity that depends only on the shape of \(R_0\) and characterizes how good \(Q_\Lambda\) is.

Assume \(R\) is a \(k\)-dimensional polytope with finite volume \(V(R)\) with centroid at the origin, i.e.,

\[\min_Y \int_R \|x - y\|^2 \, dx = \int_R \|x\|^2 \, dx\]

**Definition:** The dimensionless normalized second moment of \(R\) is

\[G(R) \triangleq \frac{1}{k} \frac{1}{V(R)^{1+2/k}} \int_R \|x\|^2 \, dx\]
For $\alpha > 0$ and $R \subset \mathbb{R}^k$, let $\alpha R = \{\alpha x : x \in R\}$.

**Claim:** $G(R)$ is scale-invariant; i.e., for any $\alpha > 0$,

$$G(\alpha R) = G(R).$$

**Proof:** Using the change of variable $y = \alpha^{-1}x$,

$$\int_{\alpha R} \|x\|^2 \, dx = \alpha^{k+2} \int_R \|y\|^2 \, dy$$

We have $V(\alpha R) = \alpha^k V(R)$. Thus

$$V(\alpha R)^{1+2/k} = \alpha^{k+2} V(R)^{1+2/k}$$

and we get

$$\frac{1}{k} \frac{1}{V(\alpha R)^{1+2/k}} \int_{\alpha R} \|x\|^2 \, dx = \frac{1}{k} \frac{1}{V(R)^{1+2/k}} \int_R \|y\|^2 \, dy$$

**Proof sketch:** Let $X_1, X_2, \ldots$ be i.i.d. with each $X_i$ being uniform on $[0, 1]$. Assume $\Lambda \subset \mathbb{R}^k$ and $\Lambda \subset \mathbb{R}^n$.

$Q_\Lambda$ has $N \approx \frac{V([0,1]^k)}{V(R_0)} = \frac{1}{V(R_0)}$ codevectors in $[0,1]^k$.

$Q_\hat{\Lambda}$ has $\hat{N} \approx \frac{V([0,1]^n)}{V(R_0)} = \frac{1}{V(R_0)}$ codevectors in $[0,1]^n$.

The comparison is fair if the per sample rates are equal:

$$\frac{1}{k} \log_2 N = \frac{1}{n} \log_2 \hat{N} \quad \iff \quad \frac{1}{k} \log_2 \frac{1}{V(R_0)} = \frac{1}{n} \log_2 \frac{1}{V(R_0)}$$

$$\iff \quad V(R_0)^{2/k} = V(\hat{R}_0)^{2/n}$$

Then

$$\frac{1}{k} D(Q_\Lambda) \approx V(R_0)^{2/k} G(R_0) = G(R_0)$$

$$\frac{1}{n} D(Q_\hat{\Lambda}) \approx V(R_0)^{2/n} G(R_0) = G(R_0)$$

Thus the lattice with smaller $G$ has smaller per dimension distortion. $\Box$

Express $D(Q_\Lambda)$ in terms of $G(R_0)$:

$$D(Q_\Lambda) \approx \frac{1}{V(R_0)} \int_{R_0} \|x\|^2 \, dx$$

$$= k V(R_0)^{2/k} \frac{1}{k} V(R_0)^{1+2/k} \int_R \|x\|^2 \, dx$$

$$= k V(R_0)^{2/k} G(R_0)$$

Thus the per sample distortion is

$$\frac{1}{k} D(Q_\Lambda) \approx V(R_0)^{2/k} G(R_0)$$

**Claim:** Under high-resolution conditions, $G(R_0)$ is the appropriate measure for comparing lattice quantizers (even if the lattices have different dimensions).

**Examples:**

- $k = 1$: The only lattice is $\mathbb{Z}$ (and its scaled versions). We have $R_0 = [-1/2, 1/2]$, $V(R_0) = 1$, and

$$G(R_0) = \frac{1}{1^{1+2/3}} \int_{-1/2}^{1/2} x^2 \, dx = \frac{1}{12} = 0.08333\ldots$$

- $k = 2$: $G(R_0)$ is the smallest for the hexagonal lattice ($R_0$ is a regular hexagon).

$$G(R_0) = \frac{5}{36\sqrt{3}} = 0.08019$$

(Optimality is not very hard to prove.)

- $k = 3$: $G(R_0)$ is the smallest for the so-called body centered cubic lattice. $R_0$ is the regular truncated octahedron.

$$G(R_0) = \frac{19}{192 \cdot 21/3} = 0.07855$$

(Optimality proof is hard.)
Definition: Minimum normalized second moment for $k$-dimensional lattices:

$$G_k = \min_{\Lambda \subseteq \mathbb{R}^k} G(R_0)$$

Proposition 1

$G_k$ is lower bounded by the normalized second moment for a $k$-dimensional sphere:

$$G_k \geq G(S_k)$$

where $S_k = \{x \in \mathbb{R}^k : \|x\| \leq 1\}$.

Proof of Proposition: Let $R$ be any $k$-dimensional set with finite volume and rescale $R$ so that $V(R) = V(S_k) = V_k$:

$$G(R) - G(S_k) = \frac{1}{k} \frac{1}{V_k^{1+2/k}} \left( \int_R \|x\|^2 \, dx - \int_{S_k} \|x\|^2 \, dx \right)$$

We have

$$\int_R \|x\|^2 \, dx = \int_{R \cap S_k} \|x\|^2 \, dx + \int_{R \cap S_k^c} \|x\|^2 \, dx$$

Since $\|x\| > 1$ for all $x \in R \cap S_k^c$,

$$\int_{R \cap S_k^c} \|x\|^2 \, dx \geq \int_{R \cap S_k^c} 1 \, dx = V(R \cap S_k^c)$$

so that

$$\int_R \|x\|^2 \, dx \geq \int_{R \cap S_k} \|x\|^2 \, dx + V(R \cap S_k^c)$$

Remark: $G(S_k)$ can be explicitly calculated:

$$G(S_k) = \frac{1}{k} \int_{S_k} \|x\|^2 \, dx = \frac{V_k^{-2/k}}{k + 2}$$

where

$$V_k = V(S_k) = \frac{\pi^{k/2}}{\Gamma(k/2 + 1)}$$

and

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \, dx, \quad t > 0$$

is the Gamma function ($\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(1) = 1$, $\Gamma(t + 1) = t\Gamma(t)$.)

Similarly, since $\|x\| \leq 1$ if $x \in S_k \cap R^c$,

$$\int_{S_k} \|x\|^2 \, dx = \int_{S_k \cap R} \|x\|^2 \, dx + \int_{S_k \cap R^c} \|x\|^2 \, dx$$

But

$$V(R) = V(R \cap S_k) + V(R \cap S_k^c)$$

and

$$V(S_k) = V(R \cap S_k^c) + V(S_k \cap R^c)$$

so $V(S_k) = V(R)$ implies that $V(R \cap S_k^c) = V(S_k \cap R^c)$. Hence

$$\int_R \|x\|^2 \, dx \geq \int_{S_k} \|x\|^2 \, dx$$

which yields $G(R) \geq G(S_k)$. \qed

Note: equality holds iff $V(R \cap S_k^c) + V(S_k \cap R^c) = 0$.  

In general, $G_k = \min_{A \subset \mathbb{R}^k} G(R_0)$ is very hard to determine (only known for $k = 1, 2, 3$).

What happens as $k \to \infty$?

- Using Stirling’s approximation it is easy to show that
  \[
  \lim_{k \to \infty} G(S_k) = \lim_{k \to \infty} \frac{\Gamma(k/2 + 1)^2/k}{(k + 2)\pi} = \frac{1}{2\pi e} = 0.058549
  \]

- It is also true (but much harder to show) that
  \[
  \lim_{k \to \infty} G_k = \inf_{k \geq 1} G_k = \frac{1}{2\pi e}
  \]

Conclusion: In large dimension, optimal lattice cells are nearly spherical (in the second moment sense).

The conditions imply $f(x) \approx f(c_i)$ for all $x \in R_i$. Thus
\[
\frac{1}{k} D(Q) = \frac{1}{k} \sum_{i=1}^{N} \int_{R_i} \|x - c_i\|^2 f(x) \, dx
\approx \frac{1}{k} \sum_{i=1}^{N} f(c_i) \int_{R_i} \|x - c_i\|^2 \, dx
= \frac{1}{k} \sum_{i=1}^{N} f(c_i) G(R_i) V(R_i)^{1+2/k}
\]
where
\[
G(R_i) = \frac{1}{k} \frac{1}{V(R_i)^{1+2/k}} \int_{R_i} \|x - c_i\|^2 \, dx
\]
(recall that $c_i$ is the centroid of $R_i$).

Given the source $X$, investigate $\min_Q E[\|X - Q(X)\|^2]$, where the minimum is taken over all $N$-level $k$-dimensional vector quantizers.

**Assumptions:**
- $N$ is large;
- $X \sim f$, where $f$ is a smooth pdf;
- overload distortion (distortion contribution of unbounded cells) is negligible;
- the bounded quantizer cells are small;
- $c_i$ is the centroid of $R_i$ w.r.t. the uniform distribution on $R_i$.

**Further assumption:** As $N \to \infty$, the codevectors of the optimal $Q$ have a smooth point density $\lambda(x)$: For any “reasonable” $R \subset \mathbb{R}^k$,
\[
\lim_{N \to \infty} \frac{1}{N} \left( \text{# of codevectors in } R \right) = \int_{R} \lambda(x) \, dx
\]
Of course, $\lambda$ must be a pdf: $\lambda(x) \geq 0$ and
\[
\int_{\mathbb{R}^k} \lambda(x) \, dx = 1
\]
Since the cell $R_i$ contains exactly one codevector $c_i$:
\[
\frac{1}{N} = \frac{1}{N} \left( \text{# of codevectors in } R_i \right) \approx \int_{R_i} \lambda(x) \, dx \approx \lambda(c_i) V(R_i)
\]
so that
\[
V(R_i) \approx \frac{1}{N\lambda(c_i)}
\]
Using this approximation

\[
\frac{1}{k} D(Q) \approx \sum_{i=1}^{N} f(c_i) G(R_i) V(R_i)^{1+2/k}
\]

\[
\approx \sum_{i=1}^{N} f(c_i) G(R_i) \frac{V(R_i)}{(N \lambda(c_i))^{2/k}}
\]

\[
= \frac{1}{N^{2/k}} \sum_{i=1}^{N} G(R_i) f(c_i) \lambda(c_i)^{2/k} V(R_i)
\]

To obtain an analytic expression, we’ll use a conjecture that is generally accepted in the quantization literature.

**Lemma 4**

The integral

\[
\int_{\mathbb{R}^k} \frac{f(x)}{\lambda(x)^{2/k}} \, dx
\]

is minimized over all choices of the pdf \( \lambda \) if and only if

\[
\lambda(x) = \frac{f(x)^{k/(2+k)}}{\int_{\mathbb{R}^k} f(y)^{k/(2+k)} \, dy}
\]

**Proof:** Similar to the scalar optimal compressor proof. Use Hölder’s inequality with

\[
u(x) = f(x)^{k/(2+k)} \lambda(x)^{2/(2+k)}, \quad v(x) = \lambda(x)^{2/(2+k)}
\]

and \( p = \frac{2+k}{k}, \quad q = \frac{2+k}{2} \).

**Gersho’s conjecture:** For large \( N \), the quantization cells of the optimal \( k \)-dimensional VQ are (approximately) the scaled, rotated, and translated copies of \( R^*_k \), the convex polytope that tessellates \( \mathbb{R}^k \) with minimum normalized moment of inertia, i.e.,

\[
G(R_i) \approx \min_{R \text{ tessellates } \mathbb{R}^k} G(R) = G(R^*_k) = C^*_k
\]

Let \( C^*_k = G(R^*_k) \). With Gersho’s conjecture \( D(Q) \) is approximated as

\[
\frac{1}{k} D(Q) \approx \frac{G(R^*_k)}{N^{2/k}} \sum_{i=1}^{N} f(c_i) \lambda(c_i)^{2/k} V(R_i)
\]

\[
\approx \frac{C^*_k}{N^{2/k}} \int_{\mathbb{R}^k} \frac{f(x)}{\lambda(x)^{2/k}} \, dx
\]

It remains to choose \( \lambda \) to minimize the integral.

Recall the notation \( \|u\|_p = \left( \int_{\mathbb{R}^k} |u(x)|^p \, dx \right)^{1/p} \).

Thus the optimal \( \lambda \) is

\[
\lambda(x) = \frac{f(x)^{k/(2+k)}}{\int_{\mathbb{R}^k} f(y)^{k/(2+k)} \, dy} = \frac{f(x)^{k/(2+k)}}{\|f\|^{k/(2+k)}_{k/(2+k)}}
\]

Hence the minimum of the integral is

\[
\int_{\mathbb{R}^k} \frac{f(x)}{\lambda(x)^{2/k}} \, dx = \int_{\mathbb{R}^k} \frac{f(x)}{\left[ f(x)^{k/(2+k)} \right]^{2/k}} \left[ \|f\|^{k/(2+k)}_{k/(2+k)} \right]^{2/k} \, dx
\]

\[
= \|f\|^{2/(2+k)}_{k/(2+k)} \int_{\mathbb{R}^k} \frac{f(x)}{\left[ f(x)^{k/(2+k)} \right]^{2/k}} \, dx
\]

\[
= \|f\|^{2/(2+k)}_{k/(2+k)} \int_{\mathbb{R}^k} f(x)^{k/(2+k)} \, dx
\]

\[
= \|f\|^{2/(2+k)}_{k/(2+k)} \cdot \|f\|^{k/(2+k)}_{k/(2+k)}
\]

\[
= \|f\|^{k/(2+k)}_{k/(2+k)}
\]
Conclusion: For large $N$, the optimal $Q$ has distortion
\[ \frac{1}{k} D(Q) \approx N^{-2/k} C_k^* \|f\|_{k/(2+k)} \]

Let $D^*_k(N)$ denote the distortion of an optimal $N$-level quantizer. The following exact result is much harder to prove:

**Theorem 5 (Bucklew & Wise)**

Assume that $X \sim f$ and that $E[\|X\|^{2+\epsilon}] < \infty$ for some $\epsilon > 0$. Then
\[ \lim_{N \to \infty} N^{2/k} \frac{1}{k} D^*_k(N) = C_k \|f\|_{k/(2+k)} \]

for a positive constant $C_k$ that depends only on the dimension $k$.

**Remark:** The constant $C_k$ in the theorem is known only for $k = 1$ and $k = 2$:

\[ C_1 = C_1^* = \frac{1}{12} = 0.08333 \ldots, \quad C_2 = C_2^* = \frac{5}{36\sqrt{3}} = 0.08019 \]

Gersho’s conjecture can be rephrased as $C_k = C_k^*$ for all $k \geq 1$.

**Remarks:**

- Using Hölder’s inequality it can be shown that
  \[ E[\|X\|^{2+\epsilon}] = \int_{\mathbb{R}^k} \|x\|^{2+\epsilon} f(x) \, dx < \infty \quad \text{for some } \epsilon > 0 \]
  implies
  \[ \int_{\mathbb{R}^k} f(x)^{k/(2+k)} \, dx < \infty \]
  Thus the optimal $\lambda$ is well defined under the condition $E[\|X\|^{2+\epsilon}] < \infty$.

- The asymptotic formula may not hold without Bucklew & Wise’s moment condition. In fact, there exists a r.v. $X$ such that $E[\|X\|^2] < \infty$ and
  \[ \lim_{N \to \infty} N^{2/k} \frac{1}{k} D^*_k(N) = \infty \]
  (Of course, $E[\|X\|^{2+\epsilon}] = \infty$ for all $\epsilon > 0$ in this case.)

**Variable-rate vector quantization**

A variable-rate vector quantizer $Q$ consists of

- encoder $E : \mathbb{R}^k \to \mathcal{I} = \{1, \ldots, N\}$
- lossless binary prefix code $\psi : \mathcal{I} \to \{0, 1\}^*$
- decoder $D : \mathcal{I} \to \{c_1, \ldots, c_N\} \subset \mathbb{R}^k$

\[ X \xrightarrow{\mathcal{E}} i \in \mathcal{I} \xrightarrow{\psi} b \in \{0, 1\}^* \xrightarrow{\psi^{-1}} i \xrightarrow{D} \hat{X} \]

\[ \hat{X} = D(\mathcal{E}(X)) = Q(X) \]
Let \( \ell(i) = \text{length of } \psi(i) \).

The rate of \( Q \) is the average expected codeword length:

\[
r_V(Q) = \frac{1}{k} E[\ell(E(X))] = \frac{1}{k} \sum_{i=1}^{N} \ell(i) P(Q(X) = c_i)
\]

Recall that the entropy of the discrete r.v. \( Q(X) \) is defined by

\[
H(Q) = -\sum_{i=1}^{N} P(Q(X) = c_i) \log_2 P(Q(X) = c_i)
\]

Variable-rate quantizers are sometimes called \textit{entropy-constrained} quantizers.

Since \( Q(X) \) can take at most \( N \) different values, we know that \( H(Q) \leq \log_2 N \).

Thus the rates of the quantizer with fixed-rate encoding and variable-rate encoding are related as

\[
r_V(Q) = \frac{1}{k} H(Q) \leq \frac{1}{k} \log_2 N = r_F(Q)
\]

Remarks:

- The more “nonuniform” the distribution of \( Q(X) \), the more can be gained by variable-rate encoding.
- \( r_V(Q) \) may be \textit{finite} even if \( N = \infty \), while \( r_F(Q) = \infty \) in this case.

Recall the \textit{lossless coding theorem}:

\[
H(Q) \leq \min_{\psi \text{ is prefix code}} \sum_{i=1}^{N} \ell(i) P(Q(X) = c_i) < H(Q) + 1
\]

If the optimal \( \psi \) is used (e.g. the Huffman code), then the per sample rate of the variable-rate quantizer is bounded as

\[
\frac{1}{k} H(Q) \leq r_V(Q) < \frac{1}{k} H(Q) + \frac{1}{k}
\]

For convenience, we redefine \( r_V(Q) \) as

\[
r_V(Q) = \frac{1}{k} H(Q) \text{ bits per sample}
\]

### Optimal variable-rate VQ performance

- Recall Gersho’s constant

\[
C_k^* = \min_{R \text{ tessellates } \mathbb{R}^k} G(R)
\]

- Let \( Q \) be the \textit{optimal} variable-rate VQ under the rate constraint \( H(Q) \leq R \) and assume that \( R \) (and so \( H(Q) \)) is large.

- Using high-rate approximations one can show that the distortion \( Q \), as a function of its entropy, is

\[
\frac{1}{k} D(Q) \approx C_k^* 2^{-\frac{1}{2}(H(Q) - h(X))}
\]
As in the fixed-rate case, an exact result is rather hard to prove:

**Theorem 6 (Zador, Gray et al.)**

Assume that $X \sim f$ and that $E[\|X\|^2]$ and $h(X)$ are finite. Then

$$
\lim_{R \to \infty} Q: \min_{\frac{1}{k}H(Q) \leq R} \frac{1}{k} D(Q) 2^{2R} = B_k 2^{\frac{2}{k} h(X)}
$$

for a positive constant $B_k$ that depends only on the dimension $k$.

**Remark:** It is an open question what the exact value of $B_k$ is, and whether or not $B_k = C_k^*$. It is only known that

$$
B_1 = C_1 = C_1^* = \frac{1}{12}.
$$

---

**Variable-rate lattice VQ**

- The number of codevectors of a $k$-dimensional fixed-rate VQ with rate $R$ bits per sample is

$$
N = 2^{kR}
$$

- Thus for a VQ without structure, the encoding complexity (nearest neighbor search) increases exponentially with the dimension if the rate is held fixed.

- Lattice quantizers are *highly structured* and may facilitate very efficient nearest neighbor search. We will consider their variable-rate performance.

---

High-rate entropy of lattice VQ assuming smooth pdf $f$:

$$
H(Q_\lambda) = -\sum_i P(X \in R_i) \log_2 P(X \in R_i)
\approx -\sum_i f(y_i) V(R_i) \log_2 f(y_i) V(R_i)
= -\sum_i (f(y_i) \log_2 f(y_i)) V(R_i)
- \sum_i (f(y_i) \log_2 V(R_i)) V(R_i)
\approx h(X) - \int_{R^k} f(x) \log_2 V(R_0) dx
= h(X) - \log_2 V(R_0)
$$

Thus good lattices (with $G(R_0)$ close to the minimum $B_k$) make good variable-rate quantizers!
Connections with rate distortion theory

- So far, we have investigated the performance of optimal vector quantizers of a fixed dimension $k$.
- Rate distortion theory characterizes the ultimate performance limit for lossy compression with vector quantizers (block codes) as $k \to \infty$.

Assume $\{X_i\} = X_1, X_2, \ldots$ is a sequence of i.i.d. random variables with $E(X_i^2) < \infty$. Let $X$ be a generic r.v. having the same distribution as the $X_i$’s.

Recall the definition of the distortion rate function of $\{X_i\}$ w.r.t. the squared error distortion:

$$D(R) = \inf_{\hat{X}: I(X;\hat{X}) \leq R} E[(X - \hat{X})^2]$$

If $X \sim f_X(x)$, a more explicit characterization of $D(R)$ is

$$D(R) = \inf \int \int (x - \hat{x})^2 f_{X|\hat{X}}(\hat{x}|x)f_X(x) \, dx \, d\hat{x}$$

where the infimum is taken over all conditional densities $f_{\hat{X}|X}(\hat{x}|x)$ such that

$$\int \int f_{\hat{X}|X}(\hat{x}|x)f_X(x) \log_2 \frac{f_{\hat{X}|X}(\hat{x}|x)}{\int f_{\hat{X}|X}(\hat{x}|x')f_X(x') \, dx'} \, dx \, d\hat{x} \leq R$$

(The pdf $f_X(x)$ is fixed.)

Recall the properties of $D(R)$:

- $D(R) \geq 0$ and is monotone decreasing;
- $D(R)$ is convex;
- $\lim_{R \to 0} D(R) = D(0) = \sigma_X^2$;
- $\lim_{R \to \infty} D(R) = 0$.

- Introduce the optimal fixed and variable-rate $k$-dimensional VQ performance in coding the $k$-block $X^k = (X_1, \ldots, X_k)$.

**Optimal fixed-rate VQ performance:**

$$D_{k,F}(R) = \inf_{Q: r_F(Q) \leq R} \frac{1}{k} D(Q)$$

**Optimal variable-rate VQ performance:**

$$D_{k,V}(R) = \inf_{Q: r_V(Q) \leq R} \frac{1}{k} D(Q)$$

Note that

$$D_{k,V}(R) \leq D_{k,F}(R)$$
Source coding theorem

The converse rate distortion theorem can be restated as:

**Theorem 7 (Converse to the rate distortion theorem)**

For all $k \geq 1$,

$$D_{k,V}(R) \geq D(R)$$

**Remarks:**

- The theorem implies that $D_{k,F}(R) \geq D(R)$.
- The theorem says that $D(R)$ is an ultimate lower bound on the distortion of any block code operating at (fixed or variable) rate $R$.

Unfortunately, $D(R)$ is known explicitly for some special distributions only.

- We proved that if $X$ is Gaussian with variance $\sigma_X^2$, then

$$D(R) = \sigma_X^2 2^{-2R}$$

- In the general case, only bounds are known. For example, we proved the Shannon lower bound: If the differential entropy $h(X)$ is finite, then for all $R \geq 0$,

$$D(R) \geq D_{SLB}(R) = \frac{1}{2\pi e} 2^{-2(R-h(X))}$$

Compare $D(R)$ with variable-rate lattice VQ performance at high rates.

- Let $Q_A$ be the lattice VQ with optimal cell:

$$G_k = \min_{A \subset \mathbb{R}^k} G(R_0).$$

- Then

$$\frac{1}{k} D(Q_A) \approx G_k 2^{-\frac{2}{k}(H(Q_A)-h(X_k))}$$

- We have $h(X_k) = h(X_1) + \cdots + h(X_k) = kh(X)$ since the $X_i$ are i.i.d.

- Thus, with $R = \frac{1}{k} H(Q_A)$,

$$\frac{1}{k} D(Q_A) \approx G_k 2^{-2(R-h(X))}$$
Compare with the ultimate limit $D(R)$:

$$1 \leq \frac{D(Q\Lambda)}{D(R)} \leq \frac{D(Q\Lambda)}{D_{SLB}(R)} \approx \frac{G_k 2^{-2(R-h(X))}}{\frac{1}{2\pi e} 2^{-2(R-h(X))}} = G_k 2\pi e$$

$k = 1$: ($Q\Lambda$ is a uniform quantizer with entropy coding) the loss is

$$10 \log_{10}(G_k 2\pi e) = 10 \log_{10}\left(\frac{2\pi e}{12}\right) = 1.53 \text{ dB}$$

In terms of rate, this corresponds to a 0.255 bit rate loss.

$k \to \infty$:

$$\lim_{k \to \infty} \frac{D(Q\Lambda)}{D(R)} = \lim_{k \to \infty} G_k 2\pi e = 1$$

For large $k$, variable-rate lattice quantizers can perform arbitrarily close to the rate-distortion limit.