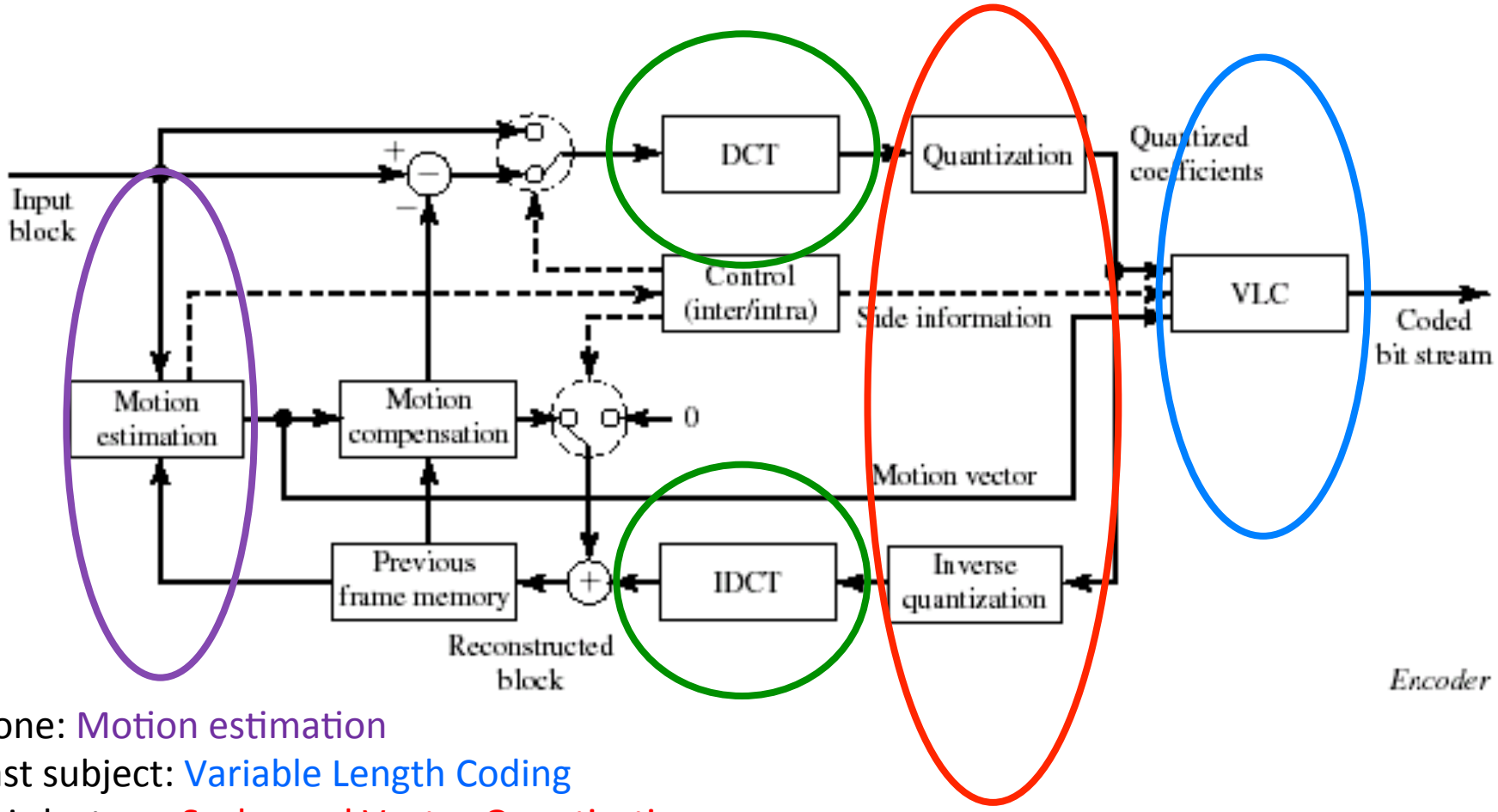
ECE 634: Digital Video Systems
Quantization : 2/9/17

Professor Amy Reibman
MSEE 356
reibman@purdue.edu

http://engineering.purdue.edu/~reibman/ece634/index.html

Encoder Block Diagram of a Typical Block-Based Video Coder
(Assuming No Intra Prediction)



Done: Motion estimation

Last subject: Variable Length Coding

This lecture: Scalar and Vector Quantization

And then: DCT, wavelet and predictive coding

Quantization

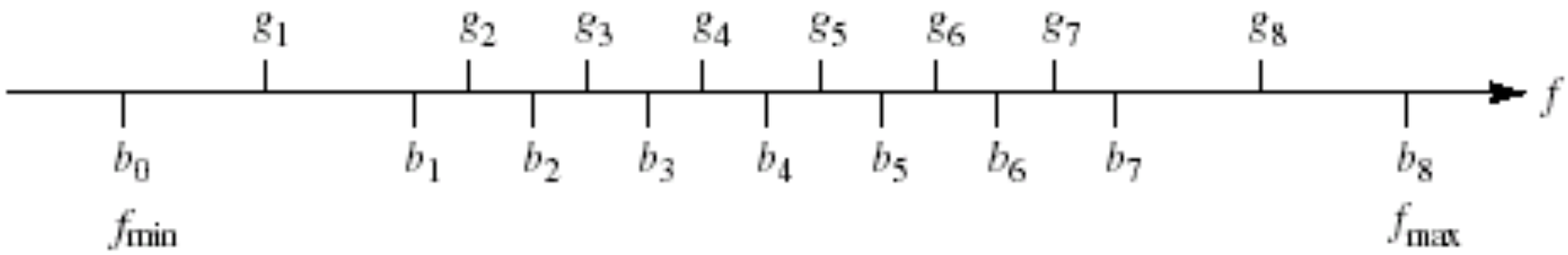
- Scalar (or memoryless) Quantization
- Vector Quantization
- Rate-distortion characterization of lossy coding
  - Operational rate distortion function
  - Rate distortion bound (lossy coding bound)

Scalar (memoryless) Quantization

- General description
- Uniform quantization
- MMSE quantizer
- Lloyd algorithm

- Quantization is inherently nonlinear

Scalar quantization partitions the line



Quantization levels: L

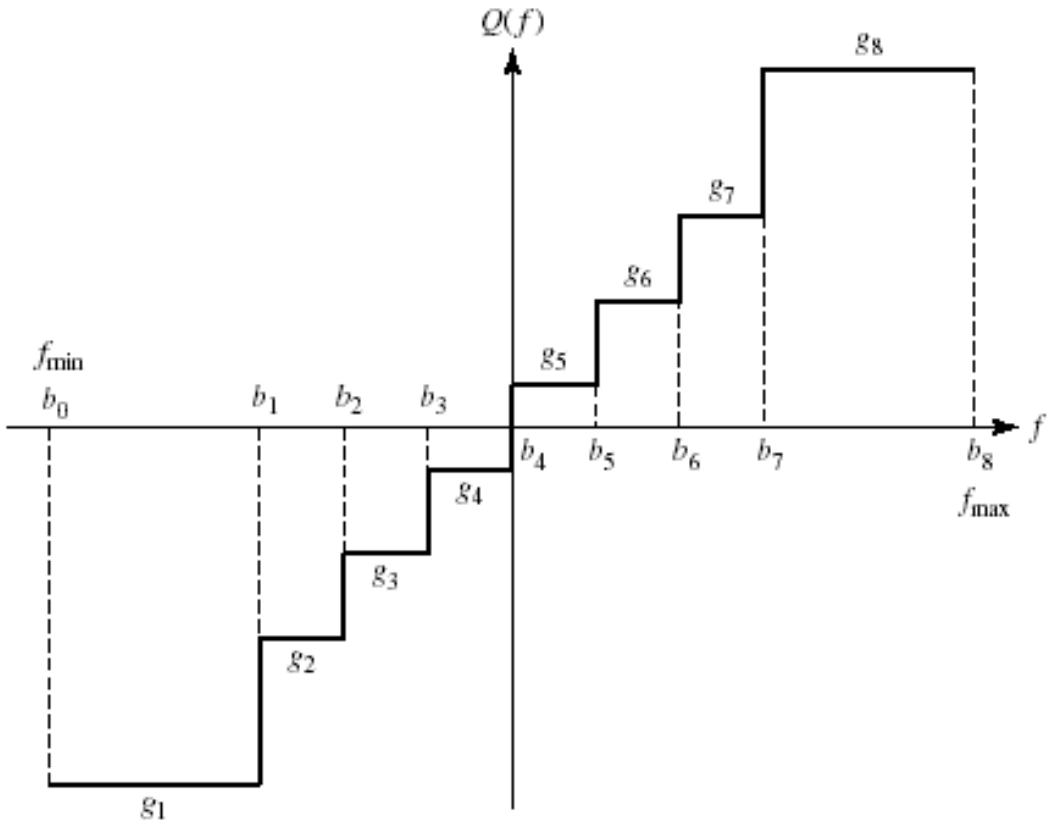
Boundary values: b_l

Partition regions: $B_l = [b_{l-1}, b_l)$

Reconstruction values: g_l

Quantizer mapping: $Q(f) = g_l$, if $f \in B_l$

Functional Representation



$$Q(f) = g_l, \text{ if } f \in B_l$$

Distortion between input and output

General distortion measure d_1(in,out):

$$D_q = E\{d_1(\mathcal{F}, Q(\mathcal{F}))\} = \int_{f\in\mathcal{B}} d_1(f, Q(f))\, p(f)\, df$$

$$= \sum_{l\in\mathcal{L}} P(\mathcal{B}_l)\, D_{q,l}$$

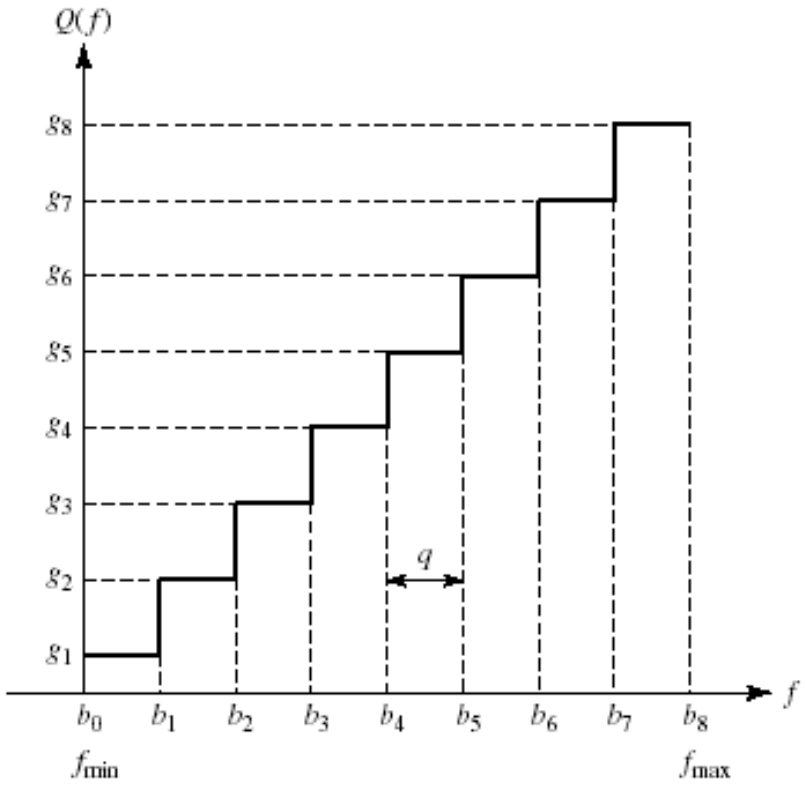
$D_{q,l}$ is expected distortion within the l-th region

$$D_{q,l} = \int_{f\in\mathcal{B}_l} d_1(f, g_l)\, p(f \mid f \in \mathcal{B}_l)\, df.$$

Mean Square Error (MSE): $d_1(f,g) = (f-g)^2$

$$\sigma_q^2 = E\{|\mathcal{F} - Q(\mathcal{F})|^2\} = \sum_{l\in\mathcal{L}} P(\mathcal{B}_l) \int_{b_{l-1}}^{b_l} (f-g_l)^2\, p(f \mid \mathcal{B}_l)\, df.$$

Uniform Quantization, MSE



$$Q(f) = \left\lfloor \frac{f - f_{\min}}{q} \right\rfloor * q + \frac{q}{2} + f_{\min},$$

Uniform source:

$$p(f) = \begin{cases} 1/B & f \in (f_{\min}, f_{\max}) \\ 0 & \text{otherwise} \end{cases}$$

$$\sigma_q^2 = \frac{q^2}{12} = \sigma_f^2 \, 2^{-2R}$$

$$\text{SNR} = 10 \log_{10} \frac{\sigma_f^2}{\sigma_q^2}$$

$$= (20 \log_{10} 2) \, R$$

$$= 6.02 R \, (\text{dB})$$

Each additional bit provides 6dB gain

Granular and overload noise

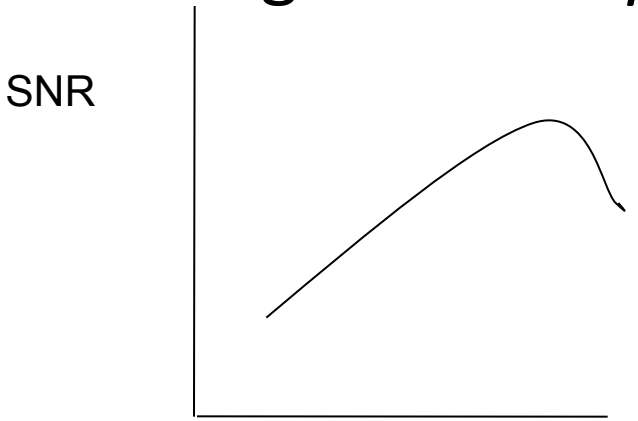
- If the quantizer does not extend far enough, we have **overload noise**
  - This can happen if there's no real lower or upper bound f_{min} and f_{max} , or if $b_0 > f_{min}$ or $b_L < f_{max}$
- The noise in the mid-range of the quantizer is called **granular noise**

- Always try to minimize overload noise
- Average distortion is the combined effect of granular and overload quantization errors

From Gersho and Gray, 1992

Uniform quantization on a non-uniform source

- High-rate approximation – many quantization levels
- Approximate pdf as uniform within any region
- Then error in any region is approximately $q^2/12$
- Average MSE is *approximately* $q^2/12$



Each additional bit provides 6dB gain, **provided** there is no overload!

From Gersho and Gray, 1992

Minimum MSE (MMSE) Quantizer

Determine b_l, g_l to minimize MSE

$$\sigma_q^2 = E\{|\mathcal{F} - Q(\mathcal{F})|^2\} = \sum_{l \in \mathcal{L}} P(\mathcal{B}_l) \int_{b_{l-1}}^{b_l} (f - g_l)^2 p(f \mid \mathcal{B}_l)\, df.$$

Setting $\frac{\partial \sigma_q^2}{b_l} = 0, \frac{\partial \sigma_q^2}{g_l} = 0$ yields:

$$b_l = \frac{g_l + g_{l+1}}{2}, \text{ or } \quad \mathcal{B}_l = \{f : d_1(f, g_l) \le d_1(f, g_{l'}), \forall l' \neq l\}.$$ (Nearest Neighbor Condition)

$$g_l = E\{\mathcal{F} \mid \mathcal{F} \in \mathcal{B}_l\} = \int_{\mathcal{B}_l} f\ p(f \mid f \in \mathcal{B}_l)\, df.$$ (Centroid Condition)

- Special case: uniform source
  - MSE optimal quantizer = Uniform quantizer

MMSE quantizer for non-uniform source

- Closed-form solutions don't always exist
- Use numerical procedures to find optimal breakpoints and reconstruction levels

- Breakpoints and reconstruction levels scale with the variance:

$$b_l = \sigma_f \tilde{b}_l + \mu_f$$

$$g_l = \sigma_f \tilde{g}_l + \mu_f$$

- Implementating quantizer: Compare all breakpoints to input f, or find reconstruction level that is closest to f

Properties of MMSE Quantizer

- Equalizes quantization error in each partition region: $D_q=P(B_l)\ D_{q,l}$
  - Recall $D_{q,l}$ is expected distortion within B_l, the l-th region, and D_q is the overall distortion
- The quantized value is an unbiased estimate of the original value: E(G)=E(F)
- The quantized value is orthogonal to the quantization error: E(GQ)=0
- The quantization processes reduces the signal variance: $\sigma^2_G=\sigma^2_F-\sigma^2_Q$

High Resolution Approximation

- For a source with arbitrary pdf, when the rate is high so that the pdf within each partition region can be approximated as flat:

$$\sigma_q^2 = \epsilon^2 \sigma_f^2 2^{-2R}$$

$$\epsilon^2 = \frac{1}{12}\left(\int_{-\infty}^{\infty} \tilde{p}(f)^{1/3}\, df\right)^3, \qquad \tilde{p}(f) = \sigma_f p(\sigma_f f)$$

Uniform source : $\varepsilon^2 = 1$

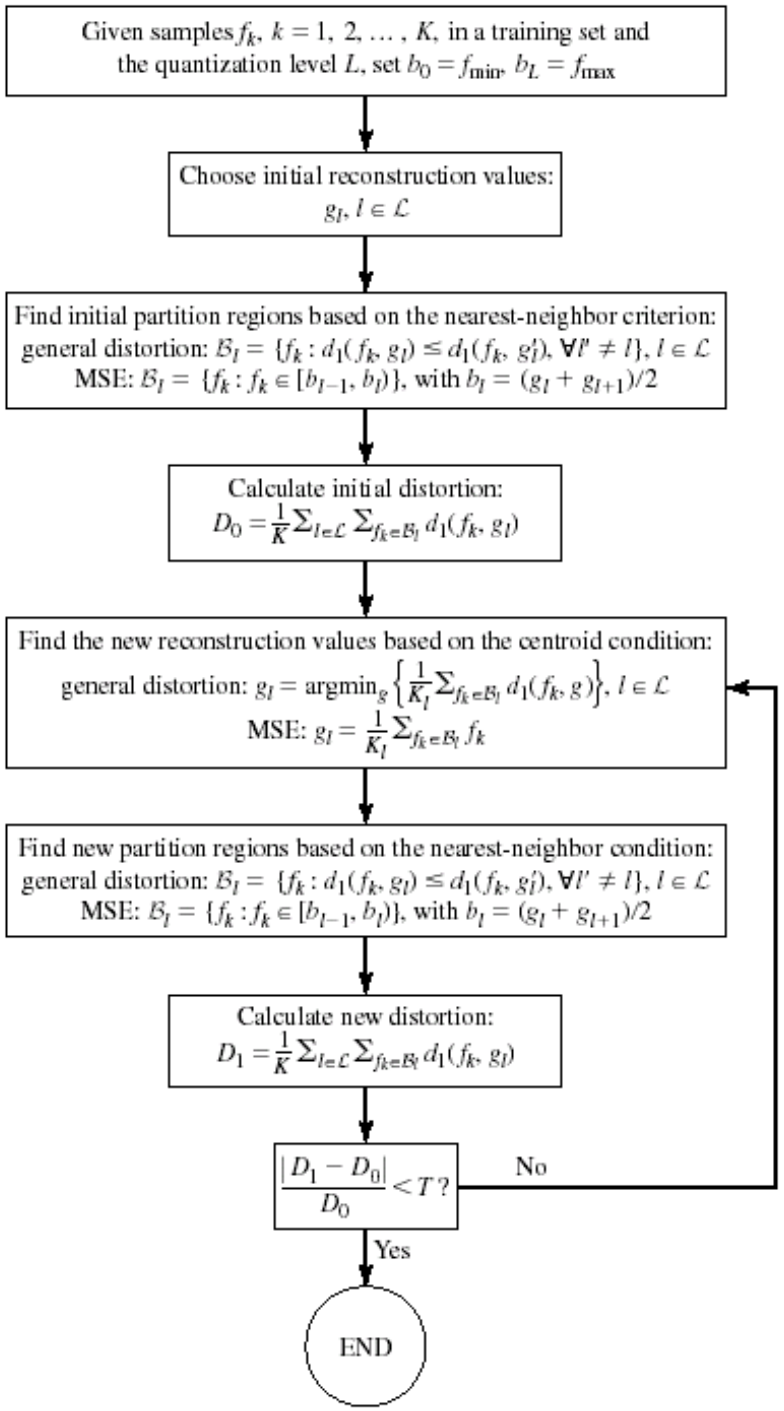
i.i.d Gaussian source : $\varepsilon^2 = 2.71$ (w/o VLC)

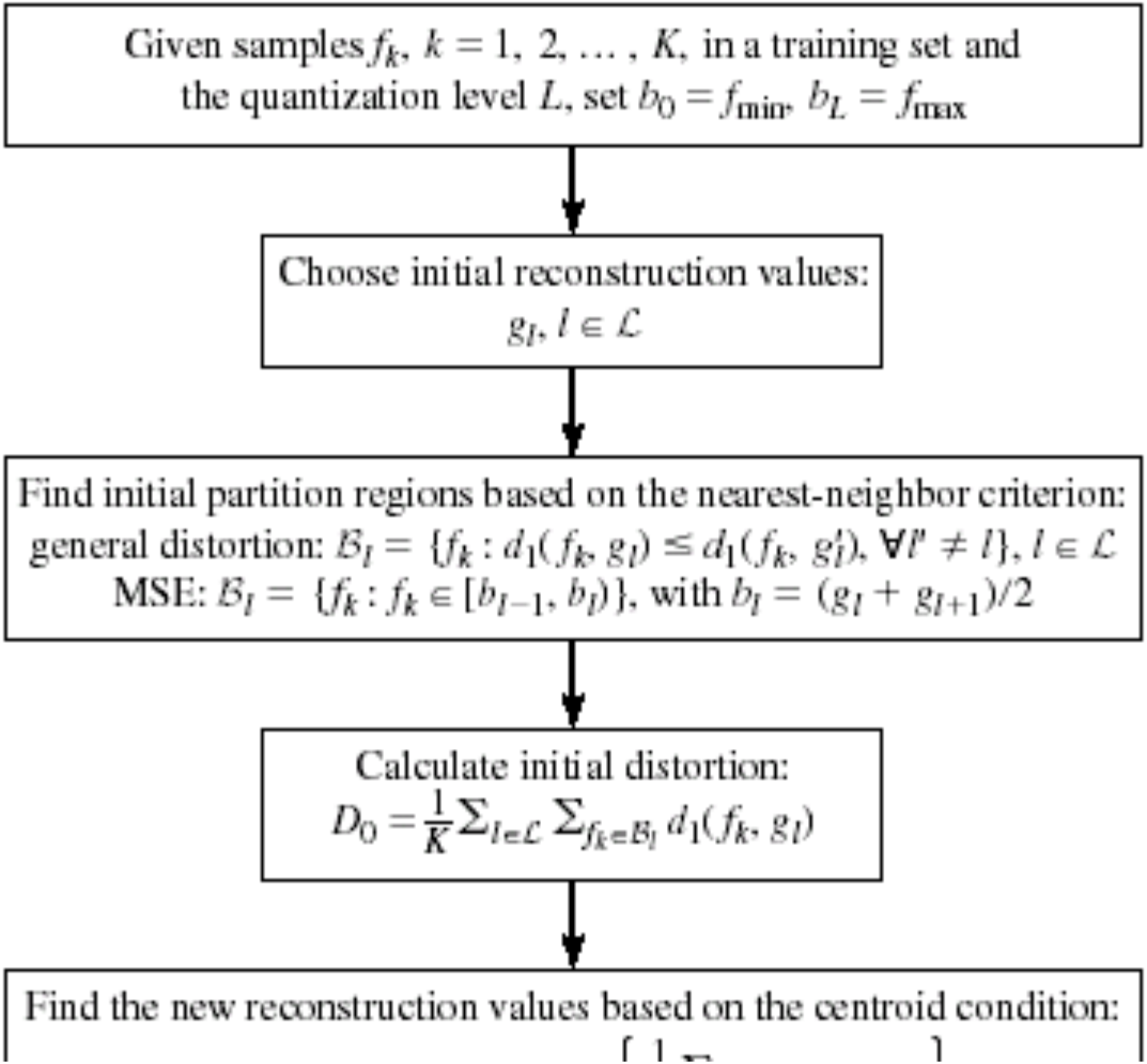
Bound for Gaussian source : $\varepsilon^2 = 1$

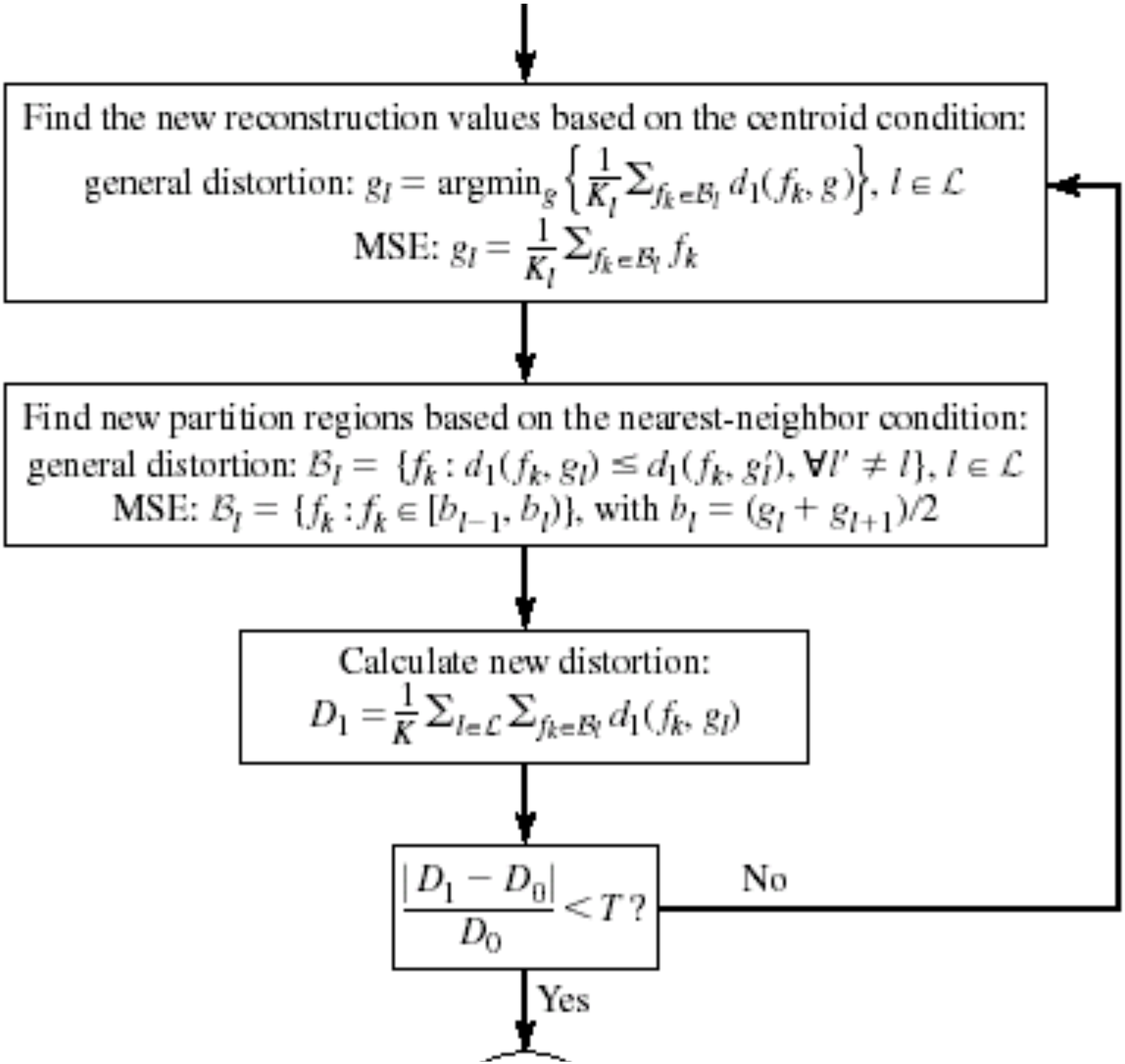
i.i.d Gaussian source with entropy coding : $\varepsilon^2 = 1.42$

Lloyd Algorithm

- Iterative algorithms for determining MMSE quantizer parameters
- Can be based on a pdf or training data
- Iterate between centroid condition and nearest neighbor condition




Given samples f_k, $k = 1, 2, \ldots, K$, in a training set and the quantization level L, set $b_0 = f_{\min}$, $b_L = f_{\max}$
Choose initial reconstruction values: g_l, $l \in \mathcal{L}$
Find initial partition regions based on the nearest-neighbor criterion:
general distortion: $\mathcal{B}_l = \{f_k : d_1(f_k, g_l) \leq d_1(f_k, g_{l'}), \forall l' \neq l\}$, $l \in \mathcal{L}$
MSE: $\mathcal{B}_l = \{f_k : f_k \in [b_{l-1}, b_l)\}$, with $b_l = (g_l + g_{l+1})/2$
Calculate initial distortion: $D_0 = \frac{1}{K} \sum_{l \in \mathcal{L}} \sum_{f_k \in \mathcal{B}_l} d_1(f_k, g_l)$
Find the new reconstruction values based on the centroid condition:


Find the new reconstruction values based on the centroid condition:
general distortion: $g_l = \text{argmin}_g \left\{ \frac{1}{K_l} \sum_{f_k \in \mathcal{B}_l} d_1(f_k, g) \right\}, l \in \mathcal{L}$
MSE: $g_l = \frac{1}{K_l} \sum_{f_k \in \mathcal{B}_l} f_k$
Find new partition regions based on the nearest-neighbor condition:
general distortion: $\mathcal{B}_l = \{f_k : d_1(f_k, g_l) \leq d_1(f_k, g_{l'}), \forall l' \neq l\}, l \in \mathcal{L}$
MSE: $\mathcal{B}_l = \{f_k : f_k \in [b_{l-1}, b_l)\}$, with $b_l = (g_l + g_{l+1})/2$
Calculate new distortion:
$D_1 = \frac{1}{K} \sum_{l \in \mathcal{L}} \sum_{f_k \in \mathcal{B}_l} d_1(f_k, g_l)$
$\frac{|D_1 - D_0|}{D_0} < T$?
No
Yes

Entropy-constrained quantization

- Does the presence of entropy coding after quantization change the quantizer design??

- Minimize the average distortion for a fixed entropy

$$\overline{R}_N = -\frac{1}{N}\sum_{l\in L} P(B_l)\log_2 P(B_l)$$

- Result: a uniform quantizer! (when the number of levels becomes large)
- Distortion in each region similar; short codewords in regions with relatively large pdf

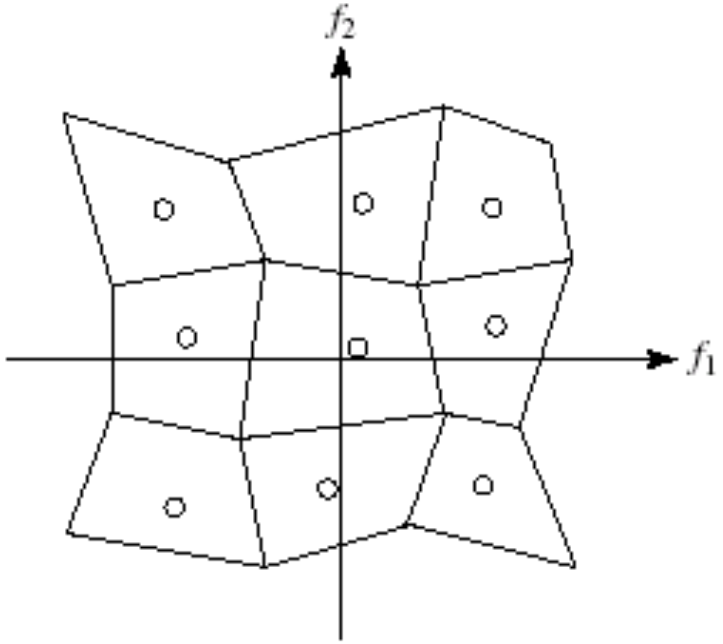
Vector Quantization

- General description
- Nearest neighbor quantizer
- MMSE quantizer
- Generalized Lloyd algorithm

Vector Quantization: General Description

- Motivation: quantize a group of samples (a vector) together, to exploit the correlation between samples
- Each sample vector is replaced by one of the representative vectors (or patterns) that often occur in the signal
- Applications:
  - Color quantization: Quantize all colors appearing in an image to *L* colors for display on a monitor that can only display *L* distinct colors at a time – Adaptive palette
  - Image quantization: Quantize every *NxN* block into one of the *L* typical patterns (obtained through training). More efficient (fewer bits per sample) with a larger block size, but the block size is limited by complexity.

VQ as Space Partition



Original vector: $\mathbf{f} \in R^N$

Quantization levels: L

Partition regions: B_l

Reconstruction vector (codeword): $\mathbf{g}_l$

Quantizer mapping: $Q(\mathbf{f}) = \mathbf{g}_l$, if $\mathbf{f} \in B_l$

Codebook: $C = \{\mathbf{g}_l, l = 1,2,...,L\}$

Bit rate: $R = \frac{1}{N}\log_2 L$

Every point in a region (B_l) is replaced by (quantized to) the point indicated by the circle ($\mathbf{g}_l$)

Distortion Measure

General distortion measure d_N(in,out):

$$D_q = E\{d_N(\mathcal{F}, Q(\mathcal{F}))\} = \int_{\mathcal{B}} p_N(\mathbf{f}) d_N(\mathbf{f}, Q(\mathbf{f}))\, d\mathbf{f}$$

$$= \sum_{l=1}^{L} P(\mathcal{B}_l)\, D_{q,l}$$

$D_{q,l}$ is expected distortion within the l-th region

$$D_{q,l} = E\{d_N(\mathcal{F}, Q(\mathcal{F})) \mid \mathcal{F} \in \mathcal{B}_l\} = \int_{\mathbf{f}\in\mathcal{B}_l} p_N(\mathbf{f} \mid \mathbf{f} \in \mathcal{B}_l) d_N(\mathbf{f}, \mathbf{g}_l)\, d\mathbf{f}.$$

MSE: $$d_N(\mathbf{f}, \mathbf{g}) = \frac{1}{N}\sum_{n=1}^{N}(f_n - g_n)^2,$$

MMSE Vector Quantizer

- Necessary conditions for MMSE
  - Nearest neighbor condition

$$\mathcal{B}_l = \{\mathbf{f} : d_N(\mathbf{f}, \mathbf{g}_l) \leq d_N(\mathbf{f}, \mathbf{g}'_l), \forall l' \neq l\}.$$

  - Generalized centroid condition:

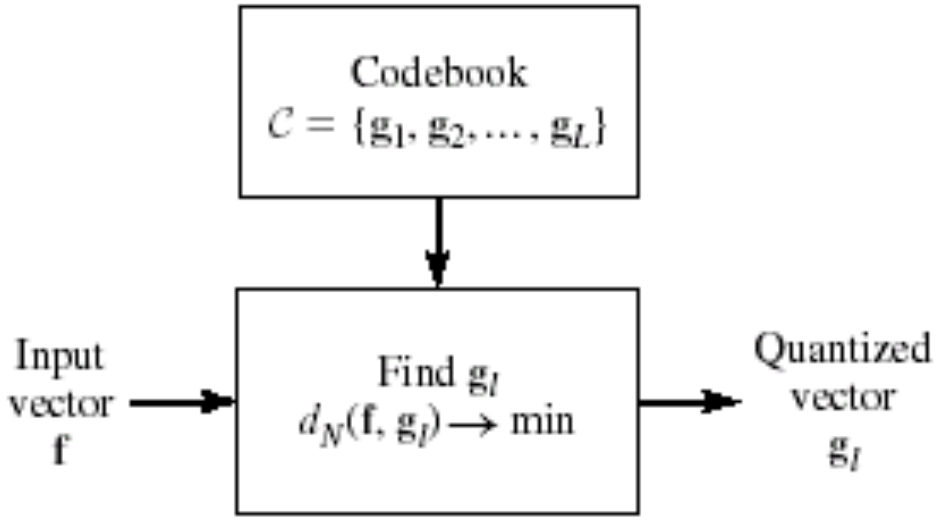
$$\mathbf{g}_l = \text{argmin}_{\mathbf{g}} E\{d_N(\mathcal{F}, \mathbf{g}) \mid \mathcal{F} \in \mathcal{B}_l\}.$$

  - MSE as distortion:

$$\mathbf{g}_l = \int_{\mathcal{B}_l} p(\mathbf{f} \mid \mathbf{f} \in \mathcal{B}_l)\mathbf{f}\, d\mathbf{f} = E\{\mathcal{F} \mid \mathcal{F} \in \mathcal{B}_l\}.$$

  - Necessary but not sufficient conditions

Nearest Neighbor (NN) Quantizer



$$\mathcal{B}_l = \{\mathbf{f} \in \mathcal{R}^N : d_N(\mathbf{f}, \mathbf{g}_l) \leq d_N(\mathbf{f}, \mathbf{g}'_l), \forall l' \neq l\}.$$

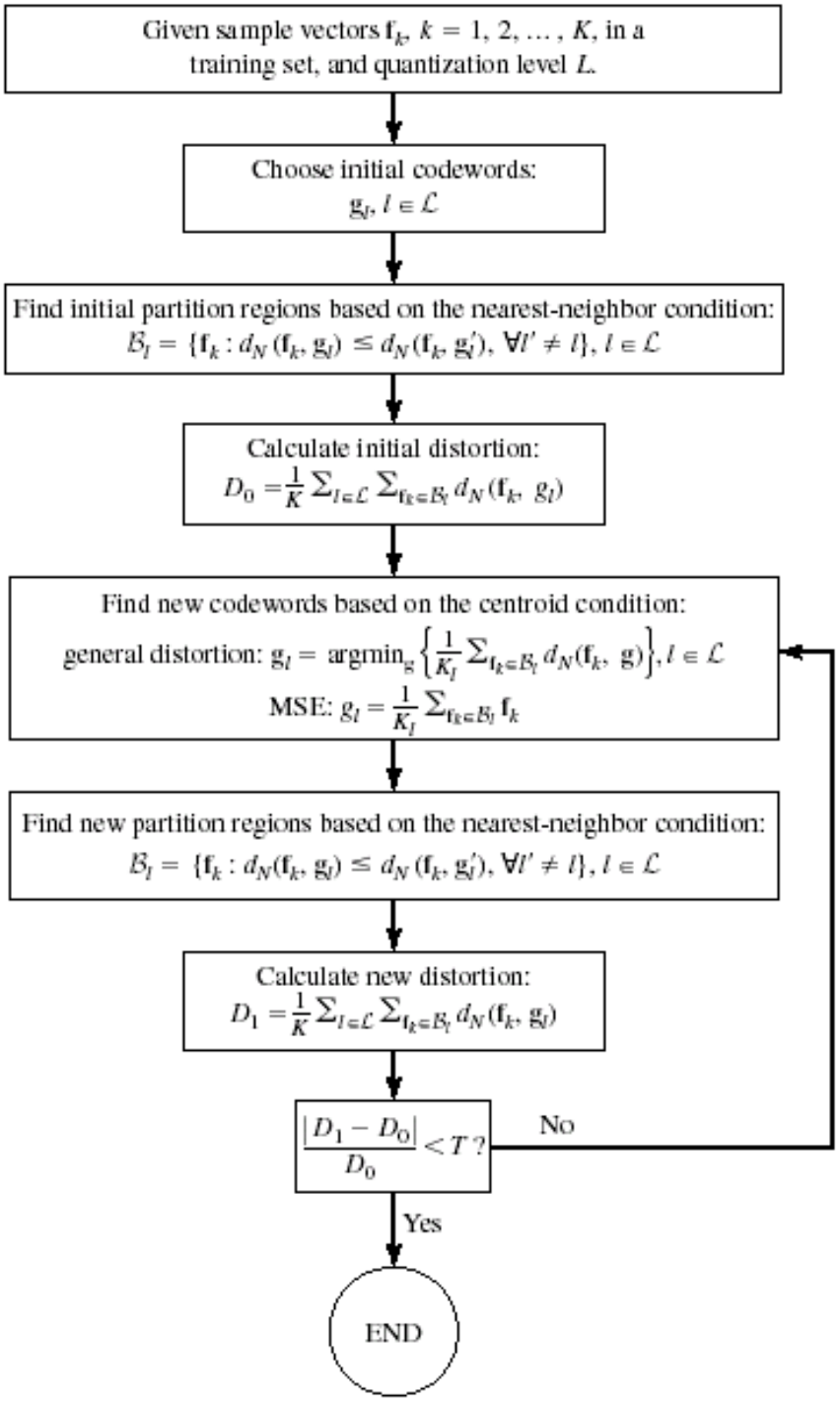
Challenge: How to determine the codebook?

Complexity of NN VQ

- Complexity analysis:
  - Must compare the input vector with all the codewords
  - Each comparison takes N operations (for each dimension)
  - Need L=2^{NR} comparisons
  - Total operation = N 2^{NR}
  - Total storage space = N 2^{NR}
  - Both computation and storage requirement increases exponentially with N!
- Example:
  - N=4x4 pixels, R=1 bpp: 16*2^16=2^20=1 Million operation/vector
  - Apply to video frames, 720*480 pels/frame, 30 fps: 2^20*(720*480/16)*30=6.8 E+11 operations/s !
  - When applied to image, block size is typically limited to <= 4x4
- Fast algorithms:
  - Structured codebook so that one can conduct binary tree search
  - Product VQ: can search subvectors separately

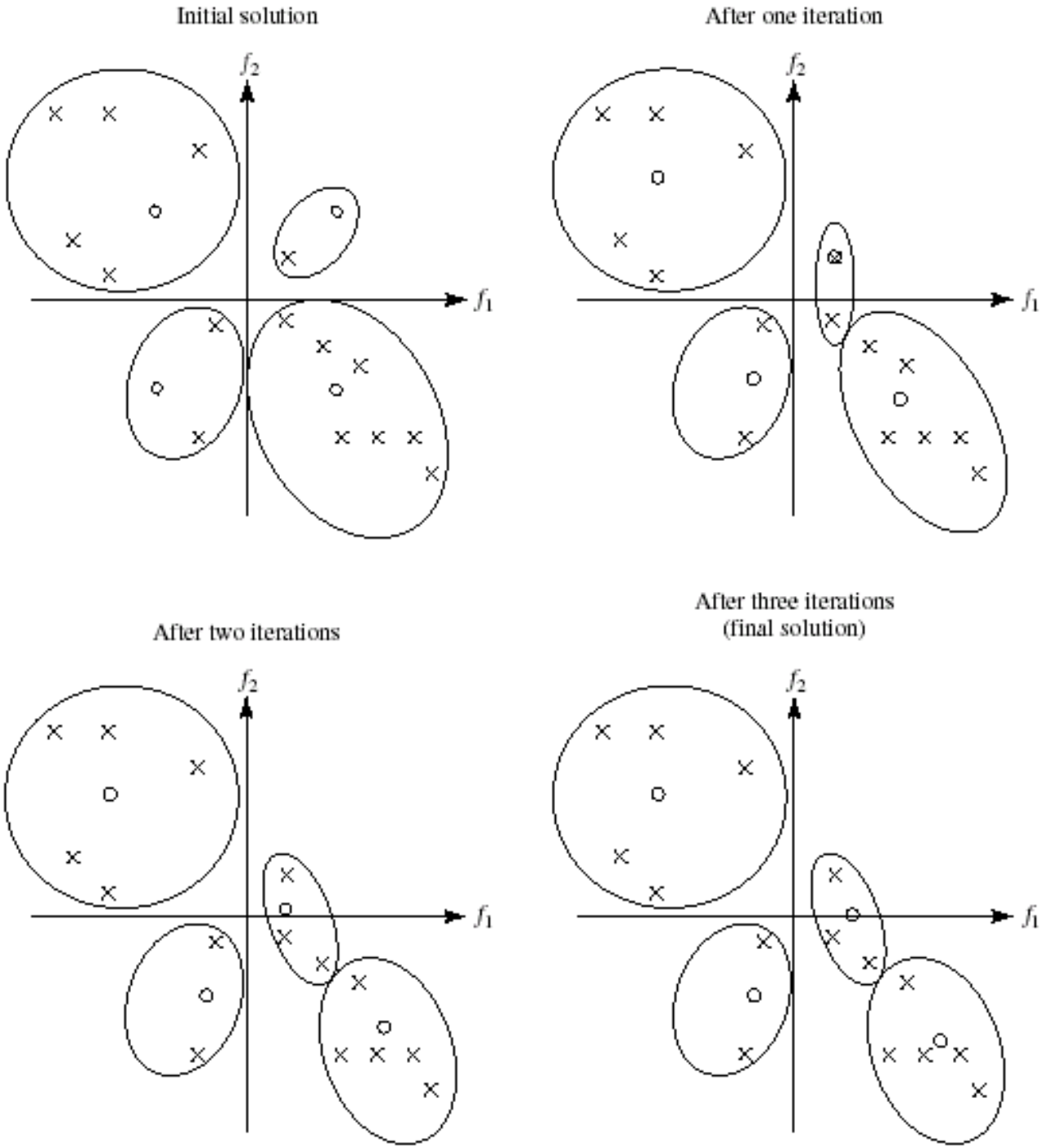
Generalized Lloyd Algorithm (LBG Algorithm)

- Start with initial codewords
- Iterate between finding best partition using NN condition, and updating codewords using centroid condition

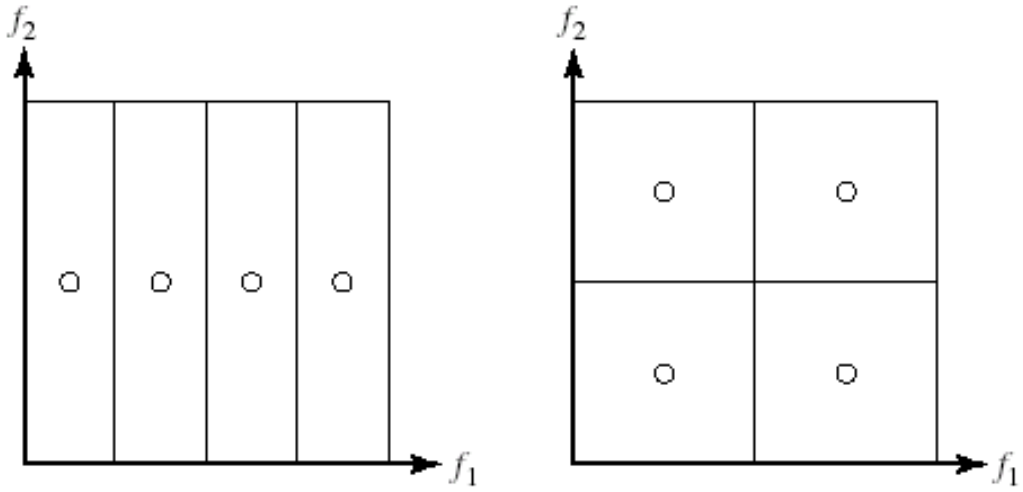


Example

x: training sample
o: reconstruction level



Caveats ☹



Both quantizers satisfy the NN and centroid condition, but the quantizer on the right is better!

NN and centroid conditions are necessary but NOT sufficient for MSE optimality!

Complexity of NN VQ

- Complexity analysis:
  - Must compare the input vector with all the codewords
  - Each comparison takes N operations (for each dimension)
  - Need L=2^{NR} comparisons
  - Total operation = N 2^{NR}
  - Total storage space = N 2^{NR}
  - Both computation and storage requirement increases exponentially with N!
- Example:
  - N=4x4 pixels, R=1 bpp: 16*2^16=2^20=1 Million operation/vector
  - Apply to video frames, 720*480 pels/frame, 30 fps: 2^20*(720*480/16)*30=6.8 E+11 operations/s !
  - When applied to image, block size is typically limited to <= 4x4
- Fast algorithms:
  - Structured codebook so that one can conduct binary tree search
  - Product VQ: can search subvectors separately

A Review of Vector Quantization

- Motivation: quantize a group of samples (a vector) together, to exploit the correlation between samples
- Each sample vector is replaced by one of the representative vectors (or patterns) that often occur in the signal
- Typically a block of 4x4 pixels
- Design is limited by ability to obtain training samples
- Implementation is limited by large number of nearest neighbor comparisons – exponential in the block size

- Transform coding+scalar quantization: a constrained vector quantizer!

Additional reading material

- Wang, Ostermann, Zhang:
  - Sec. 8.5-8.7, 8.3.2,8.3.3
- Gersho and Gray, Vector Quantization and Signal Compression, Kluwer Academic Press, 1992
  - Scalar Quantization Chapter 6&7
  - Vector Quantization Chapter 10&11

Problems to practice (1)

- Show that for an MMSE quantizer, the original random variable F, the quantized random variable G, and the quantization error Q=F-G satisfy the following relationships:
  - The quantized value is an unbiased estimate of the original value: E(G)=E(F)
  - The quantized value is orthogonal to the quantization error: E(GQ)=0
  - The quantization process reduces the signal variance: $$\sigma_G^2 = \sigma_F^2 - \sigma_Q^2$$

Problems to practice (2)

- Consider a RV F with pdf $p(f)=(\lambda/2)\exp(-\lambda|f|)$.
- A three-level quantizer is defined as
  - Find b for a given a such that the centroid condition is satisfied when the distortion measure is MSE
  - Find a for a given b such that the nearest-neighbor condition is met
  - Fin an optimal set of a,b in terms of λ such that both conditions are met. Derive the final MSE

Problems to practice (3)

- A 2-D vector quantizer has two codewords: $g_1=[1/2,1/2]^T$, $g_2=[-1/2,-1/2]^T$. Suppose that the input vectors $f=[f_1,f_2]^T$ are uniformly distributed in the square defined by $-1<f_1<1$ and $-1<f_2<1$. Illustrate the partition regions associated with the two codewords, and determine the MSE of this quantizer (it is sufficient to write the integral formula).

OPTIONAL Programming assignment

- Option 1: Write a program to perform vector quantization on a gray scale image using 4x4 pixels as a vector. You should design your codebook using all the blocks in the image as training data, using the generalized Lloyd algorithm. Then quantize the image using your codebook. You can choose the codebook size, say, L=128 or 256. Your program should work with any specified codebook size L. Observe the quality of quantized images with different L.
- Option 2: Write a program to perform color quantization on a color RGB image. Your vector dimension is now 3, containing R,G,B values. The training data are the colors of all the pixels. You should design a color palette (i.e. codebook) of size L, using generalized Lloyd algorithm, and then replace the color of each pixel by one of the color in the palette. L should be a user-selectable variable. Observe the quality of quantized images with different L.

In both cases, design your quantizer on two different images. How do the designs differ? Show results of your 2 designs on 4 different images

Rate distortion: Bounds on lossy coding performance

Mutual Information

- Mutual information between two RVs :
  - Information provided by *G* about *F*
  - Recall, *F* is the input signal, and *G* is the quantized signal

$$I(\mathcal{F};\mathcal{G}) = \sum_{f\in\mathcal{A}_f}\sum_{g\in\mathcal{A}_g} p_{\mathcal{F},\mathcal{G}}(f,g)\log_2\frac{p_{\mathcal{F},\mathcal{G}}(f,g)}{p_{\mathcal{F}}(f)\,p_{\mathcal{G}}(g)}.$$

$$I(\mathcal{F};\mathcal{G}) = H(\mathcal{F}) - H(\mathcal{F}\,|\,\mathcal{G})$$

$$I(\mathcal{F};\mathcal{G}) \le H(\mathcal{F})$$

$$I(\mathcal{F};\mathcal{G}) = H(\mathcal{F}) + H(\mathcal{G}) - H(\mathcal{F},\mathcal{G})$$

Mutual Information, vector RVs

- N-th order mutual information

$$I_N(\mathcal{F};\mathcal{G}) = \sum_{[f_1,f_2,\ldots,f_N]\in\mathcal{A}_f^N} \sum_{[g_1,g_2,\ldots,g_N]\in\mathcal{A}_g^N} p(f_1, f_2, \ldots, f_N, g_1, g_2, \ldots, g_N) \cdot \log_2 \frac{p(f_1, f_2, \ldots, f_N, g_1, g_2, \ldots, g_N)}{p(f_1, f_2, \ldots, f_N)\, p(g_1, g_2, \ldots, g_N)}$$

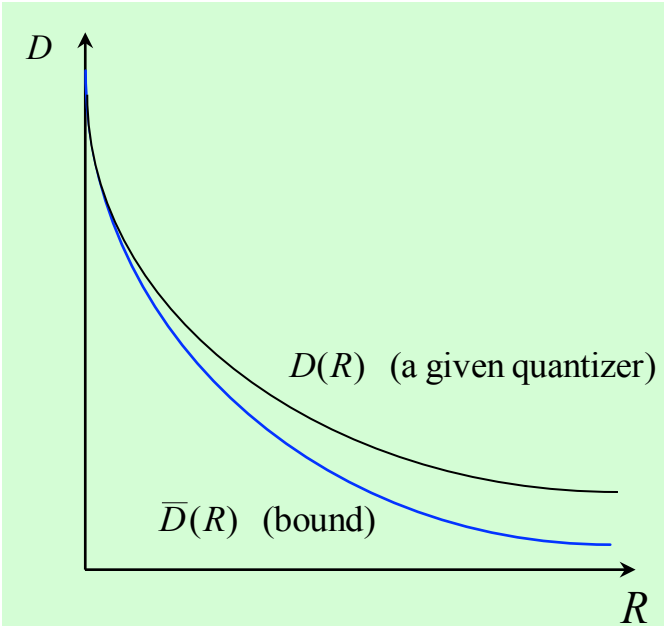
Rate-Distortion Characterization of Lossy Coding

- Rate distortion function R(D)
  - Can also use distortion-rate function D(R)
- R(D) *bound* for a source
  - The minimum rate R required to describe the source with distortion <=D

$$\bar{R}(D) = \lim_{N\to\infty} \min_{q_N(\mathbf{g}|\mathbf{f})\in Q_{D,N}} R_N(D; q_N(\mathbf{g}|\mathbf{f}))$$

$$Q_{D,N} = \{q_N(\mathbf{g}|\mathbf{f}) : E\{d_N(\mathcal{F}, \mathcal{G})\} \le D\}$$

- Operational (or empirical) rate distortion function for quantizer
  - For example, a vector quantizer reaches different points on the R(D) curve by change # regions/codewords
- R(D) optimal quantizer
  - minimize D for given R, or vice versa



Lossy Coding Bound (Shannon Lossy Coding Theorem)

$$\bar{R}(D) = \lim_{N\to\infty} \min_{q_N(\mathbf{g}|\mathbf{f})\in Q_{D,N}} R_N(D; q_N(\mathbf{g}|\mathbf{f}))$$

$$Q_{D,N} = \{q_N(\mathbf{g}|\mathbf{f}) : E\{d_N(\mathcal{F}, \mathcal{G})\} \le D\}$$

$$\bar{R}(D) = \lim_{N\to\infty} \min_{q_N(\mathbf{g}|\mathbf{f})\in Q_{D,N}} \frac{1}{N} I_N(\mathcal{F}; \mathcal{G}).$$

$I_N(\mathbf{F};\mathbf{G})$: mutual information between F and G, information provided by G about F

$Q_{D,N}$: all coding schemes (or mappings q(**g**|**f**)) that satisfy distortion criterion $d_N(f,\mathbf{g})<=D$

$$\bar{R}_L(D) \le \bar{R}(D) \le \bar{R}_G(D),$$

$$\bar{R}_L(D) = \bar{h}(\mathcal{F}) - \frac{1}{2}\log_2 2\pi e D = \frac{1}{2}\log_2 \frac{Q(\mathcal{F})}{D},$$

$h(\mathbf{F})$: differential entropy of source $\mathbf{F}$

$R_G(D)$: RD bound for Gaussian source with the same variance

i.i.d. Gaussian source requires highest bit rate!

RD bound for Gaussian source

- iid 1-D Gaussian $\bar{D}(R) = \sigma^2 2^{-2R}.$
- iid N-D Gaussian. with independent components $\bar{D}(R) = \left(\prod_n \sigma_n^2\right)^{1/N} 2^{-2R}.$
- N-D Gaussian with covariance matrix C

$$\bar{D}(R) = \left(\prod_n \lambda_n\right)^{1/N} 2^{-2R} = |\det[\mathbf{C}]|^{1/N} 2^{-2R}.$$

- Gaussian source with power spectrum $S(e^{j\omega})$
  - which is the Fourier Transform of the autocorrelation function

$$\bar{R}(D) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \log_2 \frac{S(e^{j\omega})}{D} \, d\omega.$$

Summary

- Coding system:
  - original data → model parameters → quantization → binary encoding
- Quantization:
  - Scalar quantization:
    - Uniform quantizer
    - MMSE quantizer (Nearest neighbor and centroid condition)
  - Vector quantization
    - Nearest neighbor quantizer
    - MMSE quantizer
    - Generalized Lloyd algorithm
- Rate distortion characterization of lossy coding
  - Bound on lossy coding
  - Operational RD function of practical quantizers