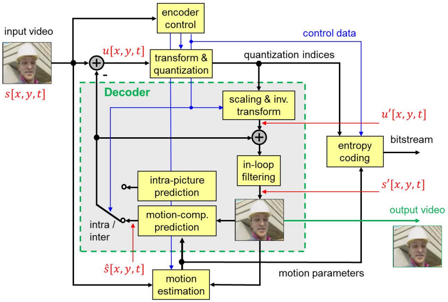
Source Coding and Compression

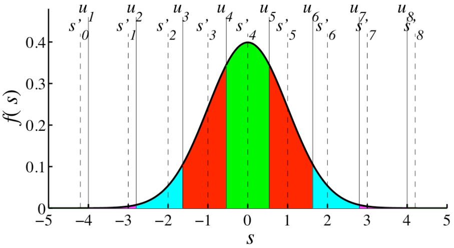
Heiko Schwarz



Contact:

Dr.-Ing. Heiko Schwarz
heiko.schwarz@hhi.fraunhofer.de

Quantization



Outline

Part I: Source Coding Fundamentals

- Probability, Random Variables and Random Processes
- Lossless Source Coding
- Rate-Distortion Theory
- **Quantization**
  - Scalar Quantization
    - Centroid Quantizer and Lloyd Quantizer
    - Entropy-Constrained Scalar Quantization
    - High-Rate Approximations for Scalar Quantizers
  - Vector Quantization
- Predictive Coding
- Transform Coding

Part II: Application in Image and Video Coding

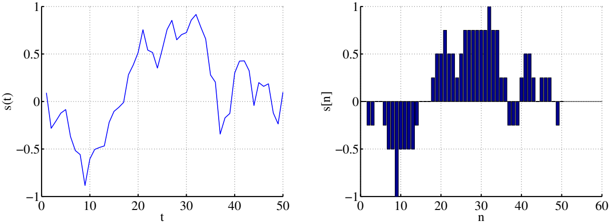
- Still Image Coding / Intra-Picture Coding
- Hybrid Video Coding (From MPEG-2 Video to H.265/HEVC)

Quantization – Introduction

- Quantization is the realization of the "lossy part" of source coding
- Typically allows for a trade-off between signal fidelity and bit rate

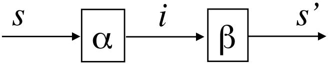
- Quantization is a functional mapping of an input point to an output point
  - the input can be discrete or continuous scalars or vectors
  - the set of obtainable output points is countable
  - less obtainable output points than input points

$\Longrightarrow$ Non-reversible loss in signal fidelity



Structure of Quantizers

- Quantizer description is split into encoder α and decoder β, between which a quantization index i is transmitted



- Adding lossless coding γ of quantization indices

- Quantization procedure
  1. Encoder α maps one or more samples of input signal s to indices i
  2. Lossless mapping γ codes the indices i into a bit stream b
  3. Channel outputs transmitted bit stream b' (error-free: $b' = b$)
  4. Inverse lossless mapping γ^{-1} reproduces quantization indices i
  5. Decoder β maps index i to one or more samples of decoded signal s'

Quantizer Mappings

- Encoder mappings α, γ have their counterparts at decoder β, γ^{-1}
- Decoder mappings must be either implemented at receiver and/or transmitted
- General case: Mapping for N-dimensional vectors

$$\boxed{Q: \ \mathbb{R}^N \rightarrow \{\boldsymbol{s'}_0, \boldsymbol{s'}_1, \cdots, \boldsymbol{s'}_{K-1}\}} \tag{344}$$

- Quantization cells: Subsets $\mathcal{C}_i$ of the N-dimensional Euclidean space $\mathbb{R}^N$

$$\mathcal{C}_i = \left\{ \boldsymbol{s} \in \mathbb{R}^N : \ Q(\boldsymbol{s}) = \boldsymbol{s'}_i \right\} \tag{345}$$

- Quantization cells $\mathcal{C}_i$ form partition of the N-dimensional Euclidean space $\mathbb{R}^N$

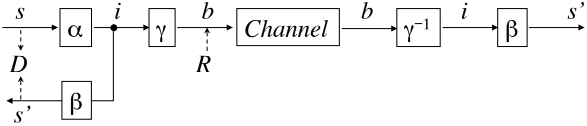
$$\bigcup_{i=0}^{K-1} \mathcal{C}_i = \mathbb{R}^N \qquad \text{with} \qquad \forall i \neq j : \ \ \mathcal{C}_i \cap \mathcal{C}_j = \emptyset \tag{346}$$

- Specify quantization mapping

$$\boxed{Q(\boldsymbol{s}) = \boldsymbol{s'}_i \qquad \forall \boldsymbol{s} \in \mathcal{C}_i} \tag{347}$$

Performance of Quantizers

- Encoder mapping $\alpha : \mathbb{R}^N \rightarrow \mathcal{I}$ introduces distortion



- Assume random process $\{\boldsymbol{S}_n\}$ to be stationary: Distortion and rate

$$\boxed{D = E\{d_N(\boldsymbol{S}_n, Q(\boldsymbol{S}_n))\} = \frac{1}{N}\sum_{i=0}^{K-1}\int_{\mathcal{C}_i} d_N(\boldsymbol{s}, Q(\boldsymbol{s}))\ f_{\boldsymbol{S}}(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}} \tag{348}$$

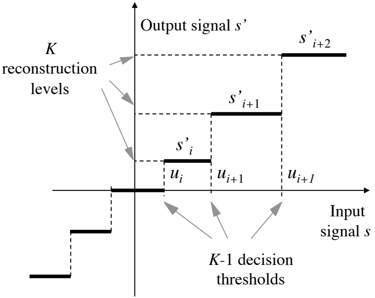
$$\boxed{R = \frac{1}{N}E\{\,|\gamma(\,Q(\boldsymbol{S}_n)\,)|\,\} = \frac{1}{N}\sum_{i=0}^{N-1} p_i \cdot |\gamma(\boldsymbol{s}'_i)| = \frac{1}{N}\sum_{i=0}^{N-1} p_i \cdot \ell_i} \tag{349}$$

where $|\gamma(\boldsymbol{s}'_i)|$ denotes codeword length ℓ_i and p_i denotes the pmf for $\boldsymbol{s}'_i$

$$p_i = p(\boldsymbol{s}'_i) = \int_{\mathcal{C}_i} f_{\boldsymbol{S}}(\boldsymbol{s})\ \mathrm{d}\boldsymbol{s} \tag{350}$$

Scalar Quantization

- Input/output function of a scalar quantizer



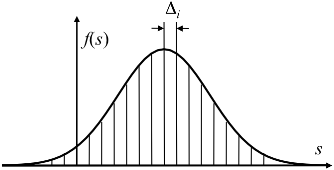
- A scalar (one-dimensional) quantizer is a mapping

$$\boxed{Q : \mathbb{R} \rightarrow \{s'_0, s'_1, \ldots, s'_{K-1}\}} \tag{351}$$

- Quantization cells $\mathcal{C}_i = [u_i, u_{i+1})$ with $u_0 = -\infty$ and $u_K = \infty$
- Step size for reconstruction level i is denoted as $\Delta_i = u_{i+1} - u_i$

Performance of Scalar Quantizers

- Scalar quantization of an amplitude-continuous random variable S can be viewed as a discretization of its continuous pdf $f(s)$



- Average MSE distortion is given as

$$\boxed{D = E\{d_1(S, Q(S))\} = E\{d_1(S, S')\} = \sum_{i=0}^{K-1} \int_{u_i}^{u_{i+1}} (s - s_i')^2 \cdot f(s)\, \mathrm{d}s} \qquad (352)$$

- Average rate is given by the expectation value of the codeword length

$$\boxed{R = E\{|\gamma(Q(S))|\} = \sum_{i=0}^{N-1} p_i \cdot |\gamma(s_i')| = \sum_{i=0}^{N-1} p_i \cdot \ell_i} \qquad (353)$$

- Goal of design: Optimize mappings α (i.e. u_i), β (i.e. s_i'), and γ

Scalar Quantization with Fixed-Length Codes

- Consider restriction on lossless mapping γ:
  $\Longrightarrow$ Assign codeword of same length to all quantization indices
- Quantizer of size K:
  $\Longrightarrow$ Codeword length must be greater than or equal to $\lceil \log_2 K \rceil$
- If K is not a power of 2, quantizer requires the same minimum codeword length as a quantizer of size $K' = 2^{\lceil \log_2 K \rceil}$
- Since $K < K'$, quantizer of size K' can achieve a smaller distortion
- Define rate according to

$$R = \log_2 K, \tag{354}$$

  while only considering quantizer sizes K that represent integer powers of 2

Simplest Case: Pulse-Code-Modulation (PCM)

- PCM: Uniform mappings α and β
  - All quantization intervals have same size Δ
  - Reconstruction values s_i' lie in the middle of the intervals
- PCM for random processes with amplitude range $[s_{\min}, s_{\max}]$

$$A = s_{\max} - s_{\min} \qquad \Longrightarrow \qquad \Delta = \frac{A}{K} = A \cdot 2^{-R} \tag{355}$$

- Quantization mapping

$$Q(s) = \operatorname{round}\left(\frac{s - s_{\min}}{\Delta} + 0.5 \right) \cdot \Delta + s_{\min} \tag{356}$$

- Example: Uniform distribution $f(s) = \frac{1}{A}$ for $-\frac{A}{2} \le s \le \frac{A}{2}$

$$D = \sum_{i=0}^{K-1} \int_{s_{\min}+i\Delta}^{s_{\min}+(i+1)\Delta} \frac{1}{A} \left(s - s_{\min} - \left(i + \frac{1}{2} \right) \cdot \Delta \right)^2 \, \mathrm{d}s \tag{357}$$

- Resulting operational rate distortion function

$$D_{\text{PCM,uniform}}(R) = \frac{A^2}{12} \cdot 2^{-2R} = \sigma^2 \cdot 2^{-2R} \tag{358}$$

PCM for Sources with Infinite Support

- In general, interval limits u_i can be chosen as

$$u_0 = -\infty,\;\; u_K = \infty,\;\; u_{i+1} - u_i = \Delta \;\text{ for }\; 1 \le i \le K-1 \qquad (359)$$

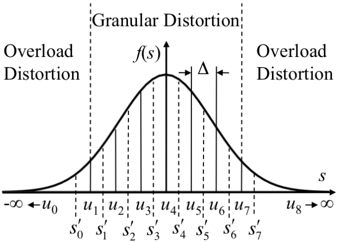
- Symmetric pdfs: Reconstruction symbols s_i with $0 \le i < K$ and interval boundaries u_i with $0 < i < K$

$$s'_i = (i - \frac{K-1}{2}) \cdot \Delta \qquad u_i = \left(i - \frac{K}{2}\right) \cdot \Delta \qquad (360)$$

- Distortion D is split into granular distortion D_G and overload distortion D_O

$$D(\Delta) = D_G(\Delta) + D_O(\Delta)$$

- Optimum Δ for given rate R?



- Distortion minimization by balancing granular and overload distortion

$$\min_{\Delta} D(\Delta) = \min_{\Delta} \left[D_G(\Delta) + D_O(\Delta) \right] \qquad (361)$$

Overload and Granular Distortion

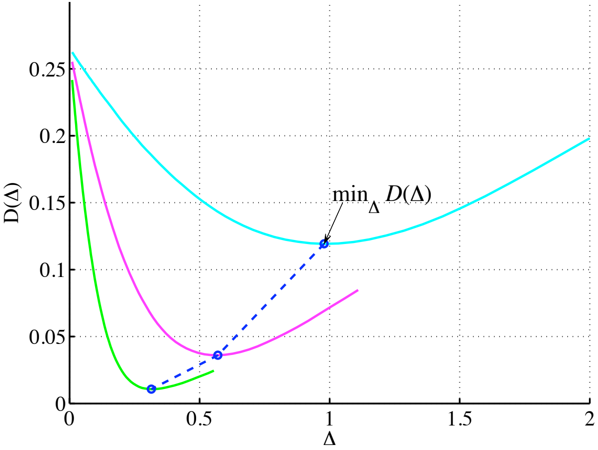
- Average distortion for PCM for sources with infinite support

$$
\begin{aligned}
D(\Delta) &= \sum_{i=0}^{K-1} \int_{u_i}^{u_{i+1}} (s - s_i')^2 \cdot f(s)\, \mathrm{d}s \\
&= \underbrace{\int_{-\infty}^{(-\frac{K}{2}+1)\Delta} (s - s_0')^2 f(s) \mathrm{d}s}_{\text{overload distortion}} \\
&\quad + \underbrace{\sum_{i=1}^{K-2} \int_{(i-\frac{K}{2})\Delta}^{(i+1-\frac{K}{2})\Delta} (s - s_i')^2 f(s)\, \mathrm{d}s}_{\text{granular distortion}} \\
&\quad + \underbrace{\int_{(\frac{K}{2}-1)\Delta}^{\infty} (s - s_{K-1}')^2 f(s)\, \mathrm{d}s}_{\text{overload distortion}}
\end{aligned}
\tag{362}
$$

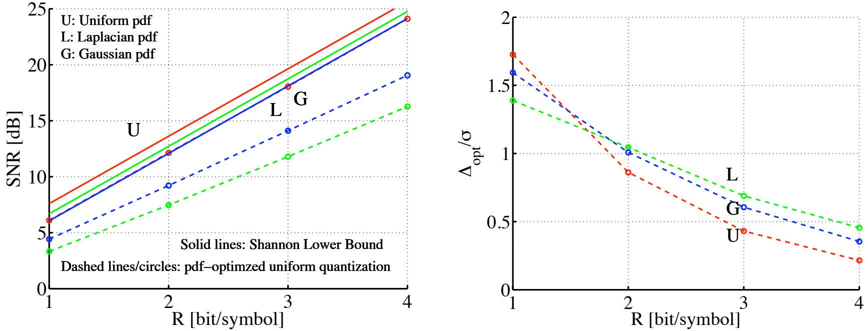
- In general: Optimum step size Δ_{opt} cannot be analytically calculated
  $\Longrightarrow$ Numerical optimization

Optimum Step Size for PCM

- Distortion $D(\Delta)$ vs. step size Δ for a Gaussian pdf with unit variance
- Cyan: $R = 2$, Magenta: $R = 3$, Green: $R = 4$ bit/sample



Numerical Optimization Results for PCM Quantization



- Numerical minimization of distortion by varying Δ
- Loss in SNR is large and increases towards higher rates
- Improvement through pdf-optimized quantizers

$\Longrightarrow$ Make quantization step sizes Δ_i variable?

$\Longrightarrow$ Modify placement of s_i' inside a quantization interval?

$\Longrightarrow$ Use variable length codes?

Optimality for Decoding Mapping: Centroid Condition

- Assume given decision thresholds and consider optimal reconstruction values
- Distortion D_i inside a quantization interval $\mathcal{C}_i$

$$D_i = \int_{u_i}^{u_{i+1}} d_1(s, s_i') \cdot f(s|s_i') \, \mathrm{d}s = E\{d_1(S, s_i') \,|\, S \in \mathcal{C}_i\} \tag{363}$$

- Probability that a source symbol falls inside quantization interval $\mathcal{C}_i$

$$p_i = \int_{u_i}^{u_{i+1}} f(s) \, \mathrm{d}s \tag{364}$$

- Average distortion

$$D = \sum_{i=0}^{K-1} p_i \cdot D_i = \sum_{i=0}^{K-1} \int_{u_i}^{u_{i+1}} d_1(s, s_i') \cdot f(s) \, \mathrm{d}s \tag{365}$$

- Since p_i does not depend on s_i', the optimality criterion is

$$\boxed{s_i'^{*} = \arg\min_{s' \in \mathbb{R}} \; E\{d_1(S, s') \,|\, S \in \mathcal{C}_i\}} \tag{366}$$

$\Longrightarrow$ **General centroid condition**

Centroid Condition for MSE Distortion

- Given a random variable X, the value of y that minimizes $E\big\{(X-y)^2\big\}$ is

$$y = E\{X\} \tag{367}$$

which can be shown by

$$\begin{aligned} E\big\{(X-y)^2\big\} &= E\big\{(X - E\{X\} + E\{X\} - y)^2\big\} \\ &= E\big\{(X - E\{X\})^2\big\} + (E\{X\} - y)^2 \\ &\geq E\big\{(X - E\{X\})^2\big\} \end{aligned} \tag{368}$$

- Consequently, given an event $\mathcal{A}$, the value y that minimizes

$$E\big\{(X-y)^2 | X \in \mathcal{A}\big\} \tag{369}$$

is

$$y = E\{X | X \in \mathcal{A}\} \tag{370}$$

Centroid Condition for MSE Distortion

- General centroid condition

$$s_i'^* = \arg\min_{s' \in \mathbb{R}} \; E\{d_1(S, s') \,|\, S \in \mathcal{C}_i\} \tag{371}$$

- MSE distortion

$$d_1(x, y) = (x - y)^2 \tag{372}$$

- The value of s_i' that minimizes the centroid condition is

$$s_i'^* = E\{S \,|\, S \in \mathcal{C}_i\} = \int_{u_i}^{u_{i+1}} s \cdot f(s|s_i') \, \mathrm{d}s = \int_{u_i}^{u_{i+1}} s \cdot \frac{f(s)}{p_i} \, \mathrm{d}s \tag{373}$$

$\Longrightarrow$ Centroid condition for MSE distortion

$$\boxed{s_i' = \frac{1}{p_i} \int_{u_i}^{u_{i+1}} s\, f(s) \, \mathrm{d}s = \frac{\int_{u_i}^{u_{i+1}} s\, f(s)\, \mathrm{d}s}{\int_{u_i}^{u_{i+1}} f(s)\, \mathrm{d}s}} \tag{374}$$

Properties of Centroid Quantizers

- Quantization does not change the mean

$$E\{S\} = \sum_i \int_{u_i}^{u_{i+1}} s\, f(s)\, \mathrm{d}s = \sum_i p_i\, s_i' = E\{S'\} = E\{Q(S)\} \tag{375}$$

- Mean of quantization error

$$E\{e(S)\} = E\{S - Q(S)\} = E\{S\} - E\{Q(S)\} = 0 \tag{376}$$

- Distortion D (2nd moment and variance of quantization error)

$$\begin{aligned} D = E\{e(S)^2\} &= \sum_i \int_{u_i}^{u_{i+1}} (s - s_i')^2\, f(s)\, \mathrm{d}s \\ &= \sum_i \left(\int_{u_i}^{u_{i+1}} s^2\, f(s)\, \mathrm{d}s - 2s_i' \int_{u_i}^{u_{i+1}} s\, f(s)\, \mathrm{d}s + s_i'^2 \int_{u_i}^{u_{i+1}} f(s)\, \mathrm{d}s \right) \\ &= \int_{-\infty}^{\infty} s^2\, f(s)\, \mathrm{d}s - \sum_i \left(2s_i' \cdot s_i' \cdot p_i - s_i'^2 \cdot p_i \right) \\ &= E\{S^2\} - E\{Q(S)^2\} \end{aligned} \tag{377}$$

$$\Longrightarrow \quad \sigma_{e(S)}^2 = \sigma_S^2 - \sigma_{Q(S)}^2 \tag{378}$$

Properties of Centroid Quantizers

- Correlation between quantizer input S and quantizer output $Q(S)$

$$\begin{aligned} E\{S \cdot Q(S)\} &= \sum_i \int_{u_i}^{u_{i+1}} s\, s'_i\, f(s)\, g(s'_i|s)\, \mathrm{d}s \\ &= \sum_i s'_i \int_{u_i}^{u_{i+1}} s\, f(s)\, \mathrm{d}s = \sum_i s'^2_i\, p_i = E\{Q(S)^2\} \end{aligned} \quad (379)$$

- Correlation between quantizer input S and quantization error $e(S)$

$$\begin{aligned} E\{S \cdot e(S)\} &= E\{S\,(S - Q(S))\} = E\{S^2\} - E\{S\,Q(S)\} \\ &= E\{e(S)^2\} + E\{Q(S)^2\} - E\{Q(S)^2\} \\ &= E\{e(S)^2\} = D \end{aligned} \quad (380)$$

- Correlation between quantizer output $Q(S)$ and quantization error $e(S)$

$$\begin{aligned} E\{Q(S) \cdot e(S)\} &= E\{Q(S)\,(S - Q(S))\} = E\{Q(S)\,S\} - E\{Q(S)^2\} \\ &= E\{Q(S)^2\} - E\{Q(S)^2\} = 0 \end{aligned} \quad (381)$$

$\Longrightarrow$ **Quantizer output and quantization error are uncorrelated**

Optimality for Encoding Mapping: Nearest Neighbor Condition

- Assume fixed-length coding and given reconstruction levels s'_i
- Choose decision thresholds u_i so that distortion D is minimized

$$D = \sum_{i=0}^{K-1} p_i \, D_i = \sum_{i=0}^{K-1} \int_{u_i}^{u_{i+1}} d_1(s, s'_i) \cdot f(s) \, \mathrm{d}s \tag{382}$$

- Each decision thresholds u_i influences only the distortions D_{i-1} and D_i of the neighboring intervals $\mathcal{C}_{i-1}$ and $\mathcal{C}_i$, respectively
- Distortion is minimized if the following condition is obeyed

$$\boxed{d_1(u_i, s'_{i-1}) = d_1(u_i, s'_i)} \tag{383}$$

- For MSE distortion, optimal decision thresholds u_i^* are given by

$$\boxed{u_i^* = \frac{s'_{i-1} + s'_i}{2}} \tag{384}$$

Lloyd Quantizer

Optimal scalar quantizer with fixed-length codes

- Do not consider entropy coding of quantization indices
- Minimize distortion for given number K of quantization intervals
- Rate can be represented by

$$R = \log_2 K \tag{385}$$

- Preferable to choose K as an integer power of 2

Necessary conditions for optimality

- General centroid condition (for reconstruction levels s'_i)

$$s'^*_i = \arg\min_{s' \in \mathbb{R}} E\{d_1(S, s') \,|\, S \in \mathcal{C}_i\} \tag{386}$$

- General nearest neighbor condition (for decision threshold u_i)

$$d_1(u_i, s'_{i-1}) = d_1(u_i, s'_i) \tag{387}$$

Lloyd Quantizer

Optimality conditions for MSE distortion

- Centroid condition

$$s'_i = \frac{\int_{u_i}^{u_{i+1}} s\, f(s)\, \mathrm{d}s}{\int_{u_i}^{u_{i+1}} f(s)\, \mathrm{d}s} \tag{388}$$

- Nearest neighbor condition (for decision threshold u_i)

$$u_i = \frac{s'_{i-1} + s'_i}{2} \tag{389}$$

Design of Lloyd quantizers

- In general, cannot be derived analytically
- Iterative algorithm consisting of
  - Optimize decision thresholds u_i given reconstruction levels s'_i
  - Optimize reconstruction levels s'_i given decision thresholds u_i
- Iterative design can be based on
  - Given probability density function (perhaps using numerical integration)
  - Sufficiently large training set for considered source

Lloyd Algorithm for a Training Set

Given is

- a sufficiently large realization $\{s_n\}$ of considered source
- the number K of reconstruction levels $\{s'_i\}$

Iterative quantizer design

1. Choose an initial set of reconstruction levels $\{s'_i\}$
2. Associate all samples of the training set $\{s_n\}$ with one of the quantization intervals $\mathcal{C}_i$ according to

$$\alpha(s_n) = \arg\min_{\forall i} \; d_1(s_n, s'_i) \qquad \text{(nearest neighbor condition)}$$

   and update the decision thresholds $\{u_i\}$ accordingly
3. Update the reconstruction levels $\{s'_i\}$ according to

$$s'_i = \arg\min_{s' \in \mathbb{R}} \; E\{d_1(S, s') \,|\, \alpha(S) = i\} \qquad \text{(centroid condition)}$$

   where the expectation value is taken over the training set
4. Repeat the previous two steps until convergence

Example: Lloyd Algorithm for a Gaussian Source

- Gaussian distribution with zero mean and unit variance

$$f(s) = \frac{1}{\sigma\sqrt{2\pi}} e^{-s^2/(2\sigma^2)} \tag{390}$$

- Draw a sufficiently large number of samples (> 10000) from $f(s)$
- Design Lloyd quantizer with rate $R = 2$ bit/symbol ($K = 4$)
- Result of Lloyd algorithm
  - Decision thresholds u_i

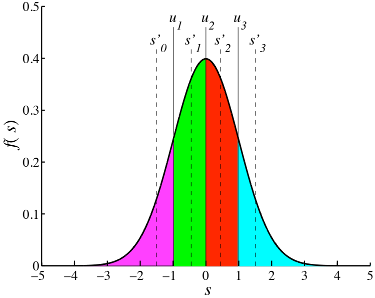
    $$u_1 = -0.98, \quad u_2 = 0, \quad u_3 = 0.98$$

  - Decoding symbols s'_i

    $$s'_0 = -1.51, \quad s'_1 = -0.45$$

    $$s'_2 = 0.45, \quad s'_3 = 1.51$$

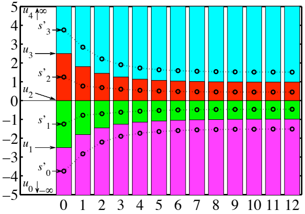
  - Minimum distortion:
    $D_F^* = 0.12 = 9.3$ dB

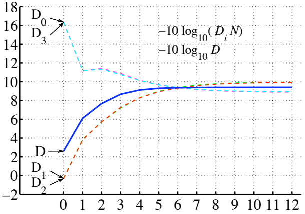


Convergence of Lloyd Algorithm for Gaussian Source Example

Initialization A:

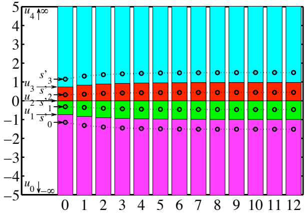
$$s'_i = -3.75 + 2.5 \cdot i$$

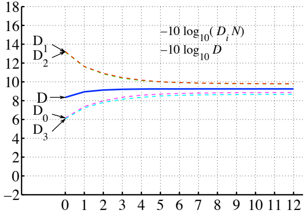




Initialization B:

$$s'_{3/0} = +/-\,1.15,\;\; s'_{2/1} = +/-\,0.32$$





- For both initializations, $(D - D_F^*)/D_F^* < 1\%$ after 6 iterations

Example: Lloyd Algorithm for a Laplacian Source

- Laplacian distribution with zero mean and unit variance

$$f(s) = \frac{1}{\sigma\sqrt{2}} e^{-|s|\sqrt{2}/\sigma} \tag{391}$$

- Draw a sufficiently large number of samples (> 10000) from $f(s)$
- Design Lloyd quantizer with rate $R = 2$ bit/symbol ($K = 4$)
- Result of Lloyd algorithm
  - Decision thresholds u_i

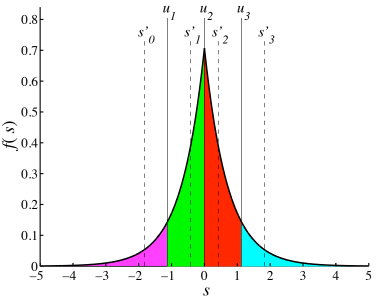
    $$u_1 = -1.13, \quad u_2 = 0, \quad u_3 = 1.13$$

  - Decoding symbols s'_i

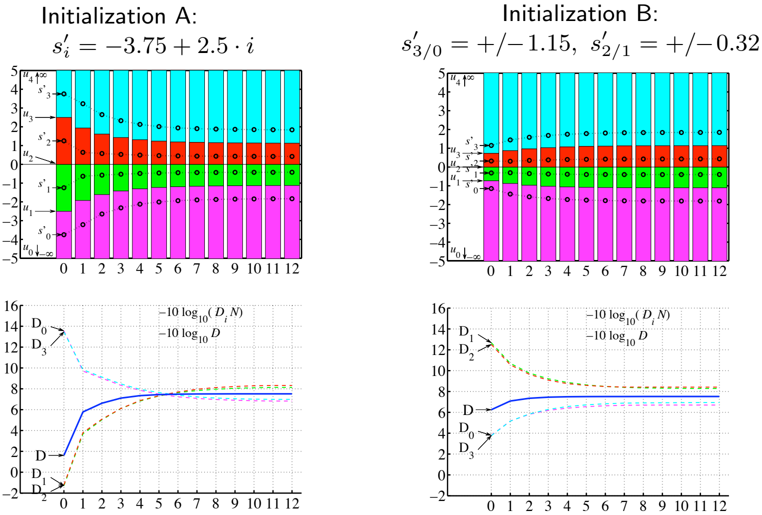
    $$s'_0 = -1.83, \quad s'_1 = -0.42$$

    $$s'_2 = 0.42, \quad s'_3 = 1.83$$

  - Minimum distortion:
    $D_F^* = 0.18 = 7.55$ dB



Convergence of Lloyd Algorithm for Laplacian Source Example



- For both initializations, $(D - D_F^*)/D_F^* < 1\%$ after 6 iterations

Entropy-Constrained Scalar Quantization (ECSQ)

- Lloyd quantizer: Minimize distortion for given number K of intervals
- Now: Consider quantizer design with variable-length coding of indices

- Average rate (without exploiting dependencies between quantization indices)

$$R = \sum_{i=0}^{N-1} p_i \cdot \ell_i \;\geq\; H(S') = -\sum_{i=0}^{K-1} p_i \log_2 p_i \tag{392}$$

with

$$p_i = \int_{u_i}^{u_{i+1}} f(s)\,\mathrm{d}s \tag{393}$$

$\Longrightarrow$ Consider entropy instead of the rate of an actual code

- Average MSE distortion

$$D = E\{d_1(S,S')\} = \sum_{i=0}^{K-1} \int_{u_i}^{u_{i+1}} (s - s'_i)^2 \cdot f(s)\,\mathrm{d}s \tag{394}$$

Joint Minimization of Rate and Distortion

- We look for solutions of constrained minimization problems

$$\min D \qquad \text{subject to} \qquad R \leq R_C \tag{395}$$

$$\text{or equivalently} \qquad \min R \qquad \text{subject to} \qquad D \leq D_C \tag{396}$$

- Instead of the constrained minimization, minimize a **Lagrangian function**

$$\boxed{J = D + \lambda \cdot R = E\{d_1(S, S')\} + \lambda \cdot E\{\ell(S')\}} \tag{397}$$

- The chosen λ corresponds to a rate constraint R_C (distortion constraint D_C)

- Minimization of J with respect to reconstruction levels s'_i is the same as the minimization of the distortion D with respect to the reconstruction levels s'_i

$\Longrightarrow$ Centroid condition still optimal for reconstruction levels (decoder $\beta(i)$)

$$\text{MSE:} \qquad s'^{*}_i = E\{S | s \in \mathcal{C}_i\} = \frac{\int_{u_i}^{u_{i+1}} s \cdot f(s)\, ds}{\int_{u_i}^{u_{i+1}} f(s)\, ds} \tag{398}$$

Necessary Conditions for Optimality

- Optimal quantizer design: Minimize Lagrange cost J for given λ

$$J = D + \lambda \cdot R = E\{d_1(S, S')\} + \lambda \cdot E\{\ell(S')\} \tag{399}$$

- Optimal reconstruction levels only depend on decision thresholds u_i

$$\boxed{s_i'^* = E\{S | s \in \mathcal{C}_i\} = \frac{\int_{u_i}^{u_{i+1}} s \cdot f(s)\, ds}{\int_{u_i}^{u_{i+1}} f(s)\, ds}} \qquad \text{(for MSE)} \tag{400}$$

- Optimal codeword lengths also depends only on decision thresholds u_i

$$\boxed{\ell_i = -\log_2 p_i = -\log_2 \left(\int_{u_i}^{u_{i+1}} f(s)\, \mathrm{d}s \right)} \tag{401}$$

- How to derive optimal decision thresholds?

Optimal Decision Thresholds

- Want to minimize J given optimal decoder β and entropy coding γ

$$J = D + \lambda \cdot R = = \sum_{\forall i} \int_{u_i}^{u_{i+1}} d_1(s, s'_i)\, f(s)\, \mathrm{d}s + \lambda \sum_{\forall i} \ell_i \int_{u_i}^{u_{i+1}} f(s)\, \mathrm{d}s \tag{402}$$

- For given reconstruction levels s'_i and codeword lengths ℓ_i:

  $\Longrightarrow$ Each decision threshold u_i only influences distortion of neighboring intervals $\mathcal{C}_{i-1}$ and $\mathcal{C}_i$

- Optimal threshold u_i:
  Each value s is assigned to the interval for which $D + \lambda R$ is minimized

$$\boxed{\alpha(s) = \arg\min_{\forall s'_i} \; d_1(s, s'_i) + \lambda\, \ell_i} \tag{403}$$

Optimal Decision Thresholds

- Lagrangian function is minimized for encoding

$$\alpha(s) = \arg\min_{\forall s'_i} \; d_1(s, s'_i) + \lambda\,\ell_i \tag{404}$$

- Optimal decision threshold u_i fulfils condition

$$d_1(u_i, s'_{i-1}) + \lambda \cdot \ell_{i-1} = d_1(u_i, s'_i) + \lambda \cdot \ell_i \tag{405}$$

- For MSE distortion, we have

$$(u_i - s'_{i-1})^2 + \lambda \cdot \ell_{i-1} = (u_i - s'_i)^2 + \lambda \cdot \ell_i \tag{406}$$

  yielding

$$\boxed{u_i^* = \frac{s'_i + s'_{i-1}}{2} + \frac{\lambda}{2} \cdot \frac{\ell_i - \ell_{i-1}}{s'_i - s'_{i-1}}} \tag{407}$$

- The decision threshold is shifted from the middle between the reconstruction values toward the reconstruction value with the longer codeword

Entropy-Constrained Lloyd Algorithm for a Training Set

Given is

- a sufficiently large realization $\{s_n\}$ of considered source
- a Lagrange parameter λ

Iterative quantizer design

1. Choose initial set of reconstruction levels $\{s'_i\}$ and codeword lengths $\{\ell_i\}$
2. Associate all samples of the traing set $\{s_n\}$ with one of the quantization intervals $\mathcal{C}_i$ according to

$$\alpha(s_n) = \arg\min_{\forall s'_i} d_1(s_n, s'_i) + \lambda\, \ell_i \tag{408}$$

   and update the decision thresholds $\{u_i\}$ accordingly
3. Update the reconstruction levels $\{s'_i\}$ according to

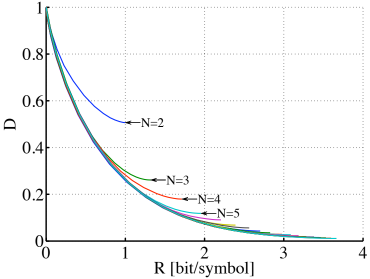
$$s'_i = \arg\min_{s' \in \mathcal{R}} E\{d_1(S, s') \mid \alpha(S) = i\} \tag{409}$$

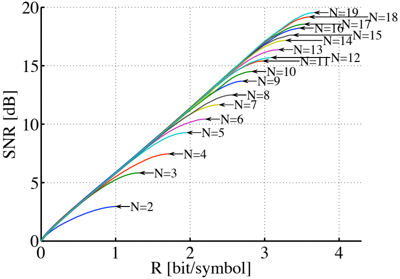
4. Update the codeword lengths ℓ_i according to

$$\ell_i = -\log_2 p_i \tag{410}$$

5. Repeat previous three steps until convergence

Number of Initial Intervals for EC Lloyd Algorithm

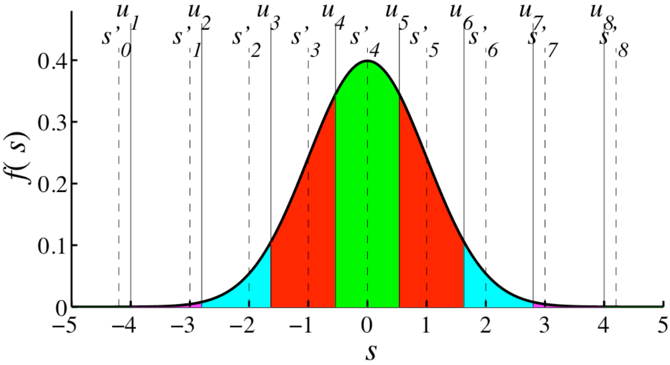




- Entropy constraint in EC Lloyd algorithm causes shift of costs
- If two level s'_i and s'_k are competing, the symbol with larger popularity has higher chance of being chosen
- Level which is not chosen further reduces its associated conditional probability
- As a consequence, symbols get "removed" and the EC Lloyd algorithm can be initialized with more symbols than the final result

Entropy-Constrained Lloyd Algorithm for Gaussian Source

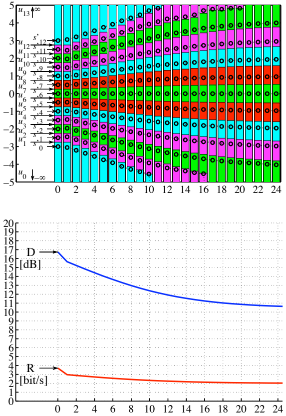
- Consider Gaussian source with zero mean and unit variance
- Design optimal entropy-constrained quantizer with rate $R = 2$ bit/symbol
- Optimum average distortion: $D_F^* = 0.09 = 10.45$ dB
- Results for optimal decision thresholds u_i and decoding symbols s'_i are



Convergence EC Lloyd Algorithm for Gaussian Source

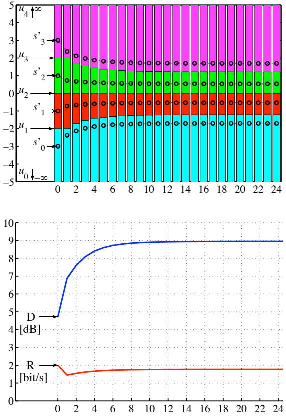
Initialization A:

$$s'_i = -3 + 0.5 \cdot i$$



Initialization B:

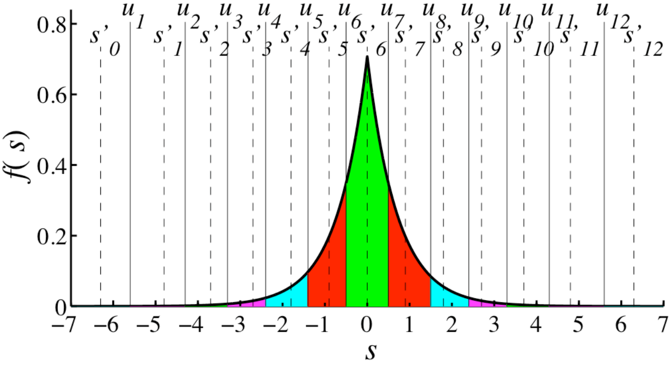
$$s'_i = -3 + 2 \cdot i$$



- For initialization A, decoding bins get discarded
- For initialization B, desired quantizer performance is not achieved

Entropy-Constrained Lloyd Algorithm for Laplacian Source

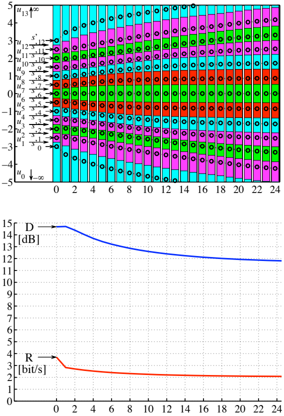
- Consider Laplacian source with zero mean and unit variance
- Design optimal entropy-constrained quantizer with rate $R = 2$ bit/symbol
- Optimum average distortion: $D_V^* = 0.07 = 11.46$ dB
- Results for optimal decision thresholds u_i and decoding symbols s_i' are



Convergence of EC Lloyd Algorithm for Laplacian Source

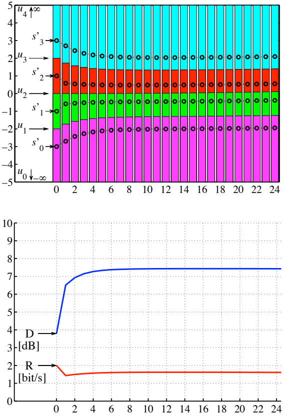
Initialization A:

$$s'_i = -3 + 0.5 \cdot i$$



Initialization B:

$$s'_i = -3 + 2 \cdot i$$



- For initialization A, faster convergence of costs than thresholds
- For initialization B, desired quantizer performance is not achieved

High-Rate Approximation for Scalar Quantizers

- Assumption: Small sizes Δ_i of quantization intervals $[u_i, u_{i+1})$
- Then: Marginal pdf $f(s)$ nearly constant inside each interval

$$f(s) \approx f(s'_i) \qquad \text{for} \qquad s \in [u_i, u_{i+1}) \tag{411}$$

- Approximation

$$p_i = \int_{u_i}^{u_{i+1}} f(s)\, ds \approx (u_{i+1} - u_i) f(s'_i) = \Delta_i \cdot f(s'_i) \tag{412}$$

- Average distortion

$$\begin{aligned} D &= E\{d(S, Q(S))\} \\ &= \sum_{i=0}^{K-1} \int_{u_i}^{u_{i+1}} (s - s'_i)^2 f(s)\, \mathrm{d}s \\ &\approx \sum_{i=0}^{K-1} f(s'_i) \int_{u_i}^{u_{i+1}} (s - s'_i)^2\, \mathrm{d}s \end{aligned} \tag{413}$$

High-Rate Approximation for Scalar Quantizers

- Average distortion

$$D \approx \sum_{i=0}^{K-1} f(s'_i) \int_{u_i}^{u_{i+1}} (s - s'_i)^2 \, ds \tag{414}$$

$$= \frac{1}{3} \sum_{i=0}^{K-1} f(s'_i) \left((u_{i+1} - s'_i)^3 - (u_i - s'_i)^3 \right) \tag{415}$$

- By differentiation with respect to s'_i, we find that for minimum distortion,

$$(u_{i+1} - s'_i)^2 = (u_i - s'_i)^2 \qquad \Longrightarrow \qquad s'_i = \frac{1}{2}(u_i + u_{i+1}) \tag{416}$$

- Average distortion at high rates

$$\boxed{D \approx \frac{1}{12} \sum_{i=0}^{K-1} f(s'_i) \, \Delta_i^3 = \frac{1}{12} \sum_{i=0}^{K-1} p_i \, \Delta_i^2} \tag{417}$$

- Average distortion at high rates for constant $\Delta = \Delta_i$

$$\boxed{D \approx \frac{\Delta^2}{12}} \tag{418}$$

High-Rate Approximation for Scalar Quantizers with FLC

- Using $\sum_{i=0}^{K-1} K^{-1} = 1$

$$D = \frac{1}{12} \sum_{i=0}^{K-1} f(s_i') \Delta_i^3 = \frac{1}{12} \left(\left(\sum_{i=0}^{K-1} f(s_i') \Delta_i^3 \right)^{\frac{1}{3}} \cdot \left(\sum_{i=0}^{K-1} \frac{1}{K} \right)^{\frac{2}{3}} \right)^3 \tag{419}$$

- Using Hölders inequality

$$\alpha + \beta = 1 \quad \Longrightarrow \quad \left(\sum_{i=a}^{b} x_i \right)^{\alpha} \cdot \left(\sum_{i=a}^{b} y_i \right)^{\beta} \geq \sum_{i=a}^{b} x_i^{\alpha} \cdot y_i^{\beta} \tag{420}$$

with equality if and only if x_i is proportional to y_i, it follows

$$D \geq \frac{1}{12} \left(\sum_{i=0}^{K-1} f(s_i')^{\frac{1}{3}} \cdot \Delta_i \cdot \left(\frac{1}{K} \right)^{\frac{2}{3}} \right)^3 = \frac{1}{12\, K^2} \left(\sum_{i=0}^{K-1} \sqrt[3]{f(s_i')}\, \Delta_i \right)^3 \tag{421}$$

- Reason for $\alpha = 1/3$: Obtain expression in which Δ_i has no exponent

High-Rate Approximations for Scalar Quantizers with FLC

- Inequality for average distortion

$$D \geq \frac{1}{12\,K^2} \left(\sum_{i=0}^{K-1} \sqrt[3]{f(s_i')}\,\Delta_i \right)^3 \tag{422}$$

becomes equality if all terms $f(s_i')\,\Delta_i^3$ are the same

- Approximation asymptotically valid for small intervals Δ_i

$$D = \frac{1}{12K^2} \left(\int_{-\infty}^{\infty} \sqrt[3]{f(s)}\,\mathrm{d}s \right)^3 \tag{423}$$

- With $1/K^2 = 2^{-\log_2 K^2} = 2^{-2R}$: Operational distortion rate function for optimal scalar quantizers with fixed-length codes

$$\boxed{D_F(R) = \sigma^2 \cdot \varepsilon_F^2 \cdot 2^{-2R} \quad \text{with} \quad \varepsilon_F^2 = \frac{1}{12\sigma^2} \left(\int_{-\infty}^{\infty} \sqrt[3]{f(s)}\,\mathrm{d}s \right)^3} \tag{424}$$

$\Longrightarrow$ Published by Panter and Dite in [Panter and Dite, 1951] and is also referred to as the **Panter and Dite formula**

Efficiency of Optimum High-Rate Quantizers with FLCs

- $D_F(R)$ for optimum high-rate scalar quantization with fixed-length codes

$$\boxed{D_F(R) = \varepsilon_F^2 \cdot \sigma^2 \cdot 2^{-2R}} \qquad (425)$$

- **Uniform pdf**:

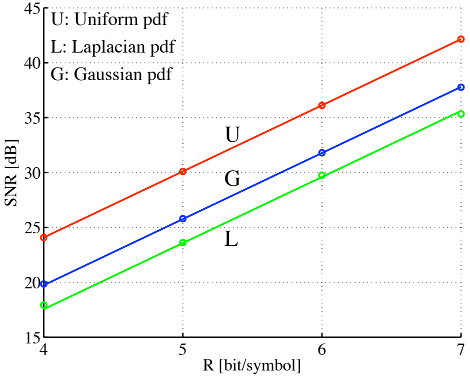
$$\varepsilon_F^2 = 1 \quad (0 \text{ dB})$$

- **Laplacian pdf**:

$$\varepsilon_F^2 = 4.5 \quad (6.53 \text{ dB})$$

- **Gaussian pdf**:

$$\varepsilon_F^2 = \frac{\sqrt{3}\pi}{2} \approx 2.721 \quad (4.35 \text{ dB})$$



High-Rate Approximation for Quantizers with VLC

- Use variable length coding for the quantizer indexes
- Again, assume pmf p_i of quantized output signal s' as $p_i = f(s'_i)\Delta_i$
- The average rate is given as

$$
\begin{aligned}
R = H(S') &= -\sum_{i=0}^{K-1} p_i \log_2 p_i = -\sum_{i=0}^{K-1} f(s'_i)\Delta_i \log_2(f(s'_i)\Delta_i) \\
&= -\sum_{i=0}^{K-1} f(s'_i) \log_2(f(s'_i)) \cdot \Delta_i - \sum_{i=0}^{K-1} f(s'_i)\Delta_i \log_2 \Delta_i \\
&\approx \underbrace{-\int f(s) \log_2 f(s)\, \mathrm{d}s}_{\text{differential entropy } h(S)} - \frac{1}{2} \sum_{i=0}^{K-1} p_i \log_2 \Delta_i^2 \\
&= h(S) - \frac{1}{2} \sum_{i=0}^{K-1} p_i \log_2 \Delta_i^2
\end{aligned}
\tag{426}
$$

High-Rate Approximation for Quantizers with VLC

- JENSEN'S inequality for convex functions $\varphi(x_i)$ such as $\varphi(x_i) = -\log_2 x_i$

$$\varphi\left(\sum_{i=0}^{K-1} a_i\, x_i\right) \leq \sum_{i=0}^{K-1} a_i\, \varphi(x_i) \qquad \text{for} \qquad \sum_{i=0}^{K-1} a_i = 1 \tag{427}$$

  with equality for constant x_i

- Jensen's inequality and the high-rate distortion approximation

$$\begin{aligned} R \;&=\; h(S) - \frac{1}{2}\sum_{i=0}^{K-1} p(s_i')\log_2 \Delta_i^2 \geq h(S) - \frac{1}{2}\log_2\left(\sum_{i=0}^{K-1} p(s_i')\Delta_i^2\right) \\ &=\; h(S) - \frac{1}{2}\log_2(12D) \end{aligned} \tag{428}$$

  with equality if and only if all $\Delta_i = \Delta$, i.e. for uniform quantization

$\Longrightarrow$ **For MSE distortion and high rates, optimal quantizers with variable length codes have uniform step sizes**

Comparison of High-Rate Distortion-Rate Functions

- **Optimum high-rate scalar quantizers with variable-length codes**

$$\boxed{D_V(R) = \frac{1}{12} \cdot 2^{2h(S)} \cdot 2^{-2R}} \tag{429}$$

is factor $\pi e/6 \approx 1.42$ or ≈ 1.53 dB from the Shannon Lower Bound (SLB)

$$D_L(R) = \frac{1}{2\pi e} \cdot 2^{2h(S)} \cdot 2^{-2R} \tag{430}$$

- Recall: **Optimum high-rate scalar quantizers with fixed-length codes**

$$D_F(R) = \frac{1}{12} \left[\int_{-\infty}^{\infty} \sqrt[3]{f(s)}\,\mathrm{d}s \right]^3 \cdot 2^{-2R} \tag{431}$$

- The $D_X(R)$ functions $(X = L, F, V)$ can be expressed in general form as

$$\boxed{D_X(R) = \varepsilon_X^2 \cdot \sigma^2 \cdot 2^{-2R}} \tag{432}$$

with ε_X^2 being a factor that depends on pdf $(f(s))$ of the source and properties of the quantizer (fixed-length vs. variable length vs. SLB)

Comparison of High-Rate Distortion-Rate Functions

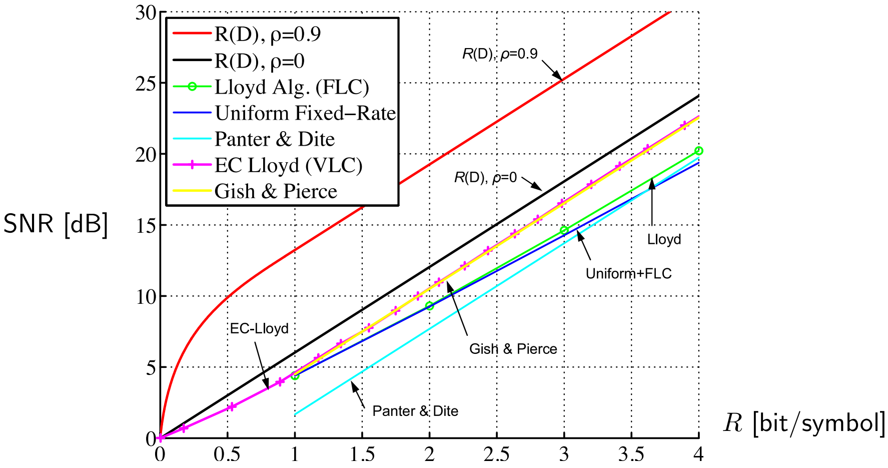
- Operational Distortion-rate function at high rates is given as

$$\boxed{D_X(R) = \varepsilon_X^2 \cdot \sigma^2 \cdot 2^{-2R}} \tag{433}$$

- Values of ε_X^2 for quantization method X

Method	*Shannon Lower Bound (SLB)*	*Panter & Dite* (Lloyd Quant. & FLC)	*Gish & Pierce* (ECSQ & VLC)
Uniform pdf	$\frac{6}{\pi e} \approx 0.7$	1 (1.53 dB to SLB)	1 (1.53 dB to SLB)
Laplacian pdf	$\frac{e}{\pi} \approx 0.86$	$\frac{9}{2} = 4.5$ (7.1 dB to SLB)	$\frac{e^2}{6} \approx 1.23$ (1.53 dB to SLB)
Gaussian pdf	1	$\frac{\sqrt{3}\pi}{2} \approx 2.72$ (4.34 dB to SLB)	$\frac{\pi e}{6} \approx 1.42$ (1.53 dB to SLB)

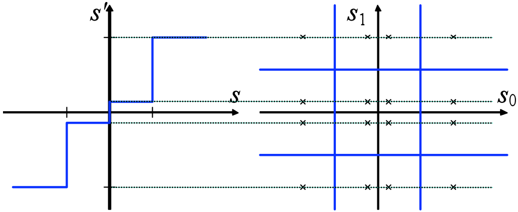
Performance of Scalar Quantizers for Gaussian Sources



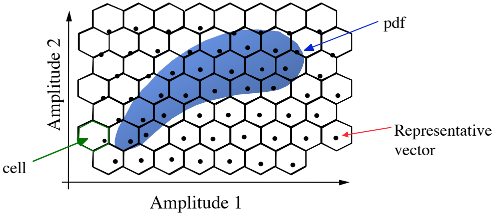
- Entropy-constrained scalar quantizer is 1.53 dB from distortion rate curve
- For sources with memory: Statistical dependencies cannot be exploited

Can We Further Improve Quantization?

- Scalar quantization: Special case of vector quantization (with $N = 1$)



- Vector quantization with $N > 1$ allows a number of new options



Vector Quantization

- Vector quantization:
  - Generalization of scalar quantization
  - Map vector of $N > 1$ samples to representative vectors
- Many models and design techniques used in vector quantization are natural generalizations of scalar quantization
- Vector quantizer Q of dimension N and size K is a mapping from a point in N-dimensional Euclidean space $\mathbb{R}^N$ into a finite set $\mathcal{C}$ containing K code vectors or code words

$$Q : \mathbb{R}^N \to \mathcal{C} \tag{434}$$

- Vector quantizer splits $\mathbb{R}^N$ into K quantization cells $\mathcal{C}_i$

$$C_i = \{\boldsymbol{s} \in \mathbb{R}^N : q(\boldsymbol{s}) = \boldsymbol{s'}\} \tag{435}$$

- The cells form a partition of $\mathbb{R}^N$

$$\bigcup_i C_i = \mathbb{R}^N \qquad \text{and} \qquad C_i \bigcap C_j = \emptyset \;\text{ for }\; i \neq j \tag{436}$$

Measuring Vector Quantizer Performance

- Average distortion for a N-dimensional vector quantizer

$$D = E\{d_N(\boldsymbol{S}, \boldsymbol{S'})\} = \int_{\boldsymbol{\mathcal{R}}^N} d_N(\boldsymbol{s}, \boldsymbol{s'}) f(\boldsymbol{s})\, d\boldsymbol{s} \tag{437}$$

- Using the partitioning of $\boldsymbol{\mathcal{R}}^N$ into cells C_i and the codebook $\boldsymbol{\mathcal{C}} = \{\boldsymbol{s}'_0, \boldsymbol{s}'_1, \ldots\}$ for a given quantizer Q

$$D = \sum_{i=0}^{K-1} \int_{C_i} d_N(\boldsymbol{s}, \boldsymbol{s}'_i) f(\boldsymbol{s})\, d\boldsymbol{s} \tag{438}$$

- For MSE distortion

$$d_N(\boldsymbol{s}, \boldsymbol{s}'_i) = \frac{1}{N} \|\boldsymbol{s} - \boldsymbol{s}'_i\| = \frac{1}{N} (\boldsymbol{s} - \boldsymbol{s}'_i)^T (\boldsymbol{s} - \boldsymbol{s}'_i) = \frac{1}{N} \sum_{n=0}^{N-1} (s_n - s'_{i,n})^2 \tag{439}$$

- Average rate (bit/scalar) for a N-dimensional vector quantizer of size K

$$R = \frac{1}{N} E\{-\log_2 p(\boldsymbol{S}'_i)\} = -\frac{1}{N} \sum_{i=0}^{K-1} p_i \log_2 p_i \tag{440}$$

The Linde-Buzo-Gray (LBG) Algorithm

Given is

- a sufficiently large realization $\{\boldsymbol{s}_n\}$ of considered source
- the number K of reconstruction vectors $\{\boldsymbol{s'}_i\}$

Iterative quantizer design (extension of Lloyd algorithm)

1. Choose an initial set of reconstruction vectors $\{\boldsymbol{s'}_i\}$
2. Associate all vectors of the training set $\{\boldsymbol{s}_n\}$ with one of the quantization cells $\mathcal{C}_i$ according to

$$\alpha(\boldsymbol{s_n}) = \arg\min_{\forall i} \; d_N(\boldsymbol{s}_n, \boldsymbol{s'}_i) \qquad \text{(nearest neighbor condition)}$$

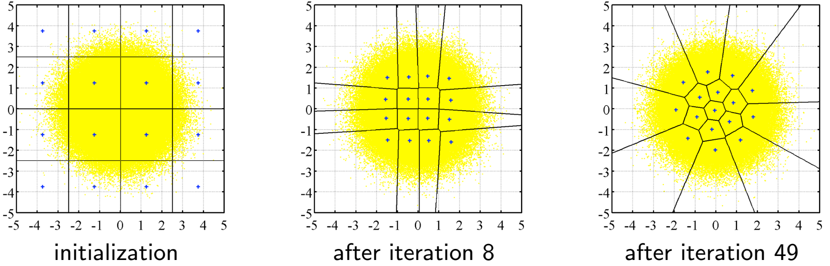
   and update the decision thresholds $\{u_i\}$ accordingly
3. Update the reconstruction vectors $\{\boldsymbol{s'}_i\}$ according to

$$\boldsymbol{s'}_i = \arg\min_{\boldsymbol{s'} \in \mathbb{R}} \; E\{d_N(\boldsymbol{S}, \boldsymbol{s'}) \,|\, \alpha(\boldsymbol{S}) = i\} \qquad \text{(centroid condition)}$$

4. Repeat the previous two steps until convergence

LBG Algorithm Result for Gaussian IID

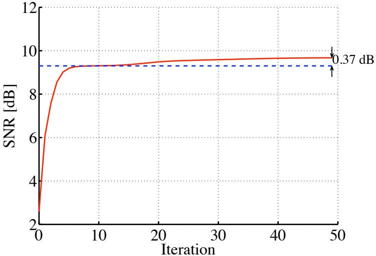
Result for dimension $N = 2$ and size $K = 16$ corresponding to $R = 2$ bit/sample



- Initialization:

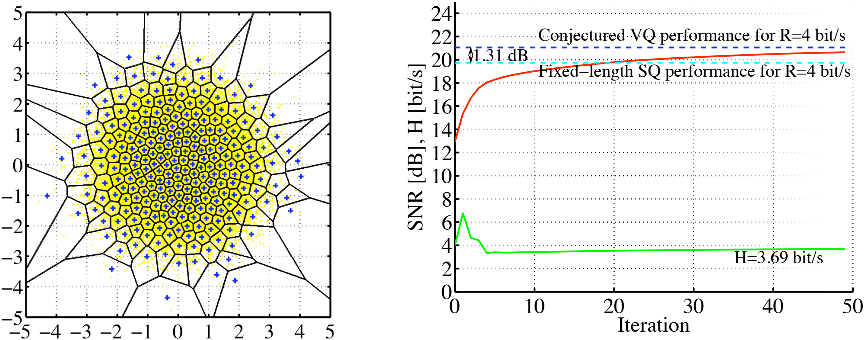
  $\boldsymbol{s'}_{i+4k} = (-3.75 + 2.5i, -3.75 + 2.5k)^T$

- After iteration 8: Same performance as in scalar case: 9.3 dB
- After iteration 49: Improvement to 9.67 dB



LBG Algorithm Result for Gaussian IID

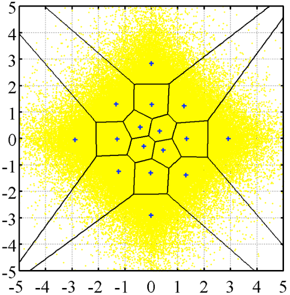
Result for dimension $N = 2$ and size $K = 256$ corresponding to $R = 4$ bit/sample

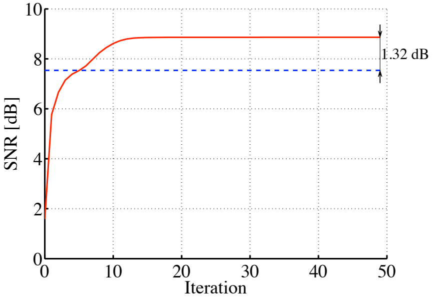


- Random initialization
- Gain around 0.9 dB for two-dimensional VQ compared to SQ with fixed-length codes resulting in 20.64 dB (of conjectured 21.05 dB)

LBG Algorithm Result for Laplacian IID

Result for dimension $N = 2$ and size $K = 16$ corresponding to $R = 2$ bit/sample





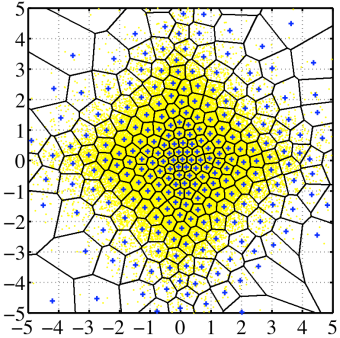
- Initialization (equal to experiment with Gaussian iid):

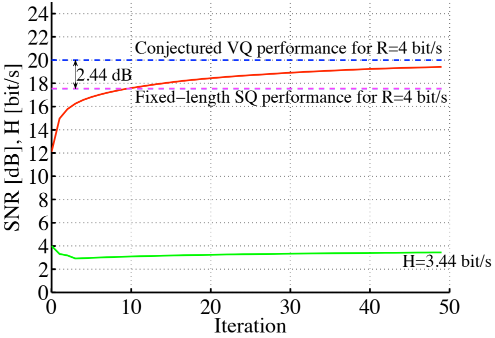
$$\boldsymbol{s'}_{i+4k} = (-3.75 + 2.5i, -3.75 + 2.5k)^T$$

- Large gain (1.32 dB) for two-dimensional VQ compared to SQ with fixed-length codes resulting in 8.87 dB

LBG Algorithm Result for Laplacian IID

Result for dimension $N = 2$ and size $K = 256$ corresponding to $R = 4$ bit/sample





- Random initialization
- Large gain (1.84 dB) for two-dimensional VQ compared to SQ with fixed-length codes resulting in 19.4 dB (of conjectured 19.99 dB)

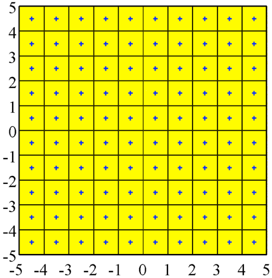
The Vector Quantizer Advantage

Gain over scalar quantization can be assigned to 3 effects

- **Space filling advantage**:
  - $\mathbb{Z}$ lattice is not most efficient sphere packing in N dimensions $(N > 1)$
  - Independent from source distribution or statistical dependencies
  - Maximum gain for $N \to \infty$: 1.53 dB
- **Shape advantage**:
  - Exploit shape of source pdf
  - Can also be exploited using entropy-constrained scalar quantization
- **Memory advantage**:
  - Exploit statistical dependencies of the source
  - Can also be exploited using predictive coding, transform coding, block entropy coding or conditional entropy coding

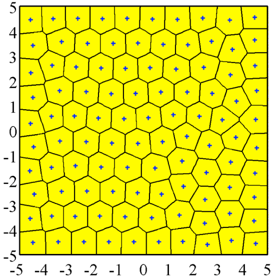
Space Filling Advantage

Consider uniform iid source with $f(s) = 1/A$ for $-A/2 \le s \le A/2$ and $A = 10$



$\Longrightarrow$

50 iterations
of LBG algorithm



- $D_U(R)$ for SQ of uniform distribution is given as $D_U(R) = \frac{A^2}{12} 2^{-2R}$; with $A = 10$ and $R = 3.32$ bit/scalar we have $D_U(R) = 19.98$ dB
- LBG algorithm converged towards 20.08 dB showing an approximate hexagonal lattice in 2D

Space-Filling Advantage: Densest Sphere Packings

Densest packings, highest kissing numbers, and approximate gain using VQ

Dim.	Densest Packing	Name	Highest Kissing Number	Approximate Gain [dB]
1	$\mathbb{Z}$	– Integer lattice	2	0
2	A_2	– Hexagonal lattice	6	0.17
3	$A_3 \simeq D_3$	– Cuboidal lattice	12	0.29
4	D_4		24	0.39
5	D_5		40	0.47
6	E_6		72	0.54
7	E_7		126	0.60
8	E_8	– Gosset lattice	240	0.66
9	Λ_9	– Laminated lattice	240	0.70
10	P_{10c}	– Non-lattice arrangement	336	0.74
12	K_{12}	– Coxeter-Todd lattice	756	0.81
16	$BW_{16} \simeq \Lambda_{16}$	– Barnes-Wall lattice	4320	0.91
24	Λ_{24}	– Leech lattice	196560	1.04
100				1.35
∞				1.53

Chou-Lookabaugh-Gray Algorithm: ECVQ

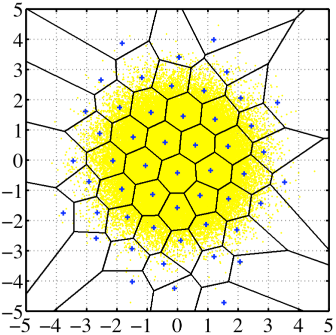
Given is

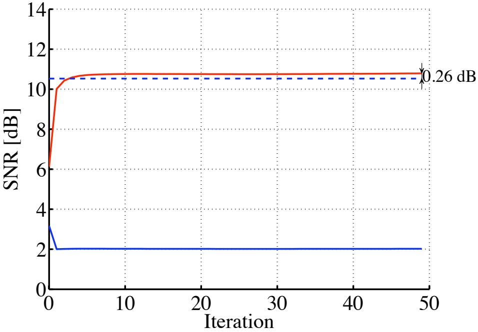
- a sufficiently large realization $\{\boldsymbol{s}_n\}$ of considered sources
- a Lagrange parameter λ

Iterative quantizer design (extension of EC Lloyd algorithm)

1. Choose initial set of reconstruction vectors $\{\boldsymbol{s}'_i\}$ and codeword lengths $\{\ell_i\}$
2. Associate all samples of the training set $\{\boldsymbol{s}_n\}$ with one of the quantization cells $\mathcal{C}_i$ according to
$$\alpha(\boldsymbol{s}_n) = \arg\min_{\forall \boldsymbol{s}'_i} \; d_N(\boldsymbol{s}_n, \boldsymbol{s}'_i) + \lambda \cdot \ell_i$$
3. Update the reconstruction vectors $\{\boldsymbol{s}'_i\}$ according to
$$\boldsymbol{s}'_i = \arg\min_{\boldsymbol{s}' \in \mathcal{R}} E\{d_N(\boldsymbol{S}, \boldsymbol{s}') \,|\, \alpha(\boldsymbol{S}) = i\}$$
4. Update the codeword lengths ℓ_i according to
$$\ell_i = -\log_2 p_i$$
5. Repeat previous three steps until convergence

Shape Advantage: Results for Gaussian IID ($N = 2$, $K = 16$)

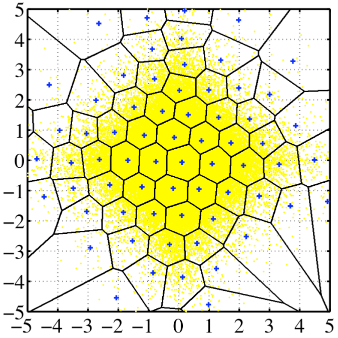


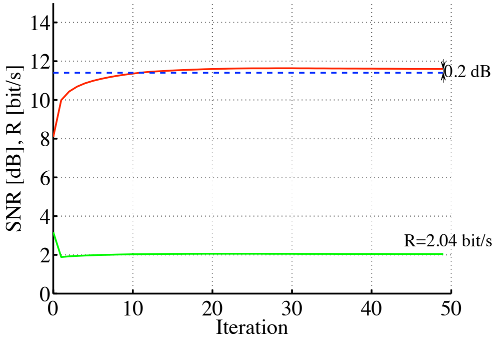


Result of CLG algorithm for Gaussian iid

- Gain of ECVQ compared to ECSQ is 0.26 dB
- Gain of VQ compared to SQ with fixed-length codes is 0.37 dB

Shape Advantage: Results for Laplace IID ($N = 2$, $K = 16$)

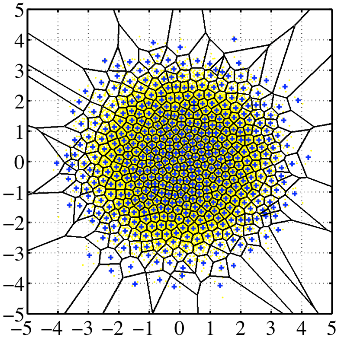


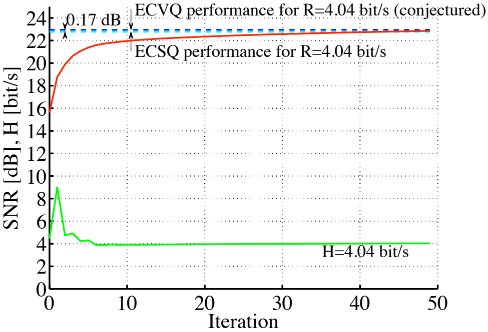


Result of CLG algorithm for Laplace iid

- Gain of ECVQ compared to ECSQ is 0.20 dB
- Gain of VQ compared to SQ with fixed-length codes is 1.32 dB

Shape Advantage: Results for Gaussian IID ($N = 2$, $K = 256$)

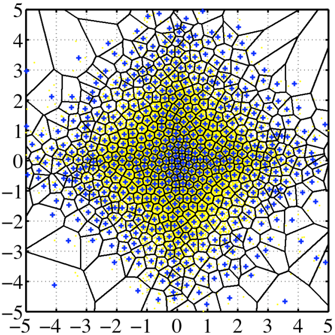


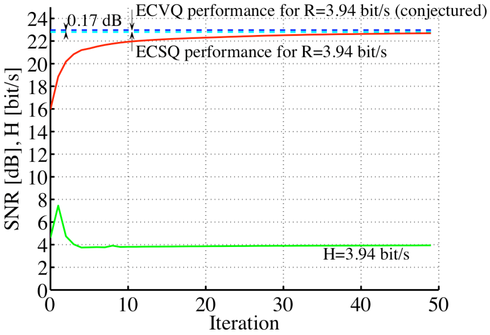


Result of CLG algorithm for Gaussian iid

- Gain of ECVQ compared to ECSQ is 0.17 dB
- Gain of VQ compared to SQ with fixed-length codes is 0.9 dB

Shape Advantage: Results for Laplace IID ($N = 2$; $K = 256$)





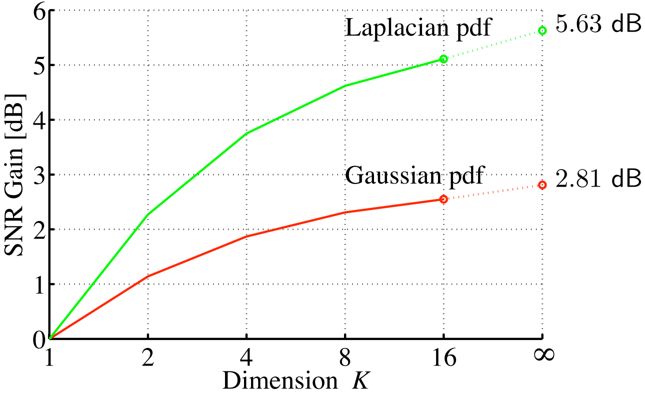
Result of CLG algorithm for 2D Laplace i.i.d.

- Gain of ECVQ compared to ECSQ is 0.17 dB
- Gain of VQ compared to SQ with fixed-length codes is 1.84 dB

$\Longrightarrow$ Entropy coding of quantization indices only leaves the space-filling gain, which is approximately 0.17 dB for $N = 2$

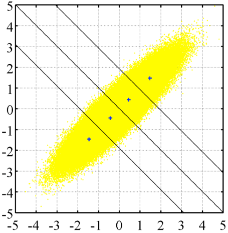
Summary on Shape Advantage

- When comparing ECSQ with ECVQ for iid sources, the gain due to $K > 1$ reduces to the space filling gain
- VQ with fixed-length codes can also exploit the gain that ECSQ shows compared to SQ with fixed-length codes



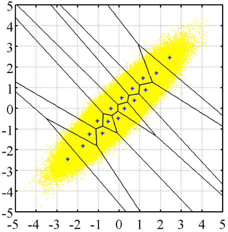
Memory Advantage: Results for Gauss-Markov with $\rho = 0.9$

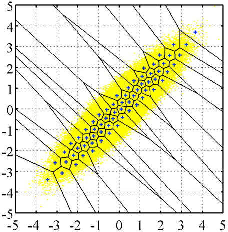
VQ results from LBG algorithm for Gauss-Markov source with correlation $\rho = 0.9$



$\Longleftarrow$ $R = 1$ bit/scalar

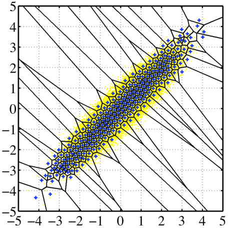
$R = 2$ bit/scalar $\Longrightarrow$





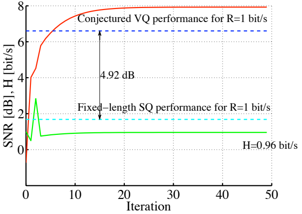
$\Longleftarrow$ $R = 3$ bit/scalar

$R = 4$ bit/scalar $\Longrightarrow$



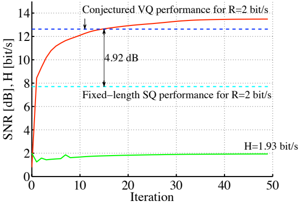
LBG algorithm has been extended by re-inserting discarded symbols s'_i using random choices

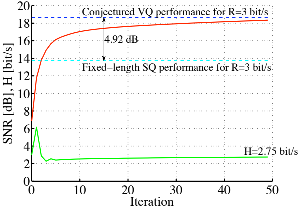
Memory Advantage: Results for Gauss-Markov with $\rho = 0.9$



$\Leftarrow$ 1 bit/scalar

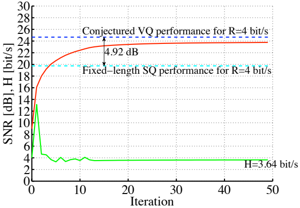
2 bit/scalar $\Rightarrow$





$\Leftarrow$ 3 bit/scalar

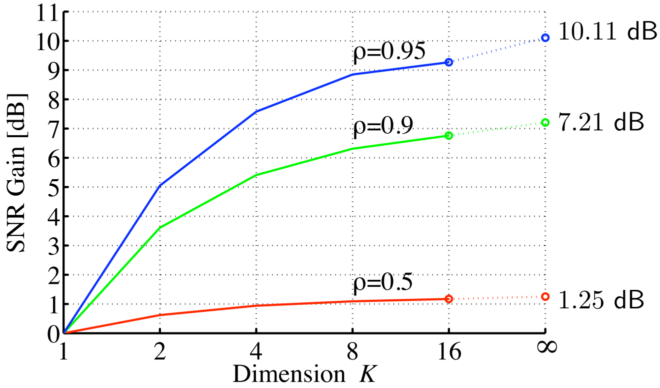
4 bit/scalar $\Rightarrow$



- Gains are additive from space-filling, shape and memory effects
- For high rates, conjectured VQ performance is approached

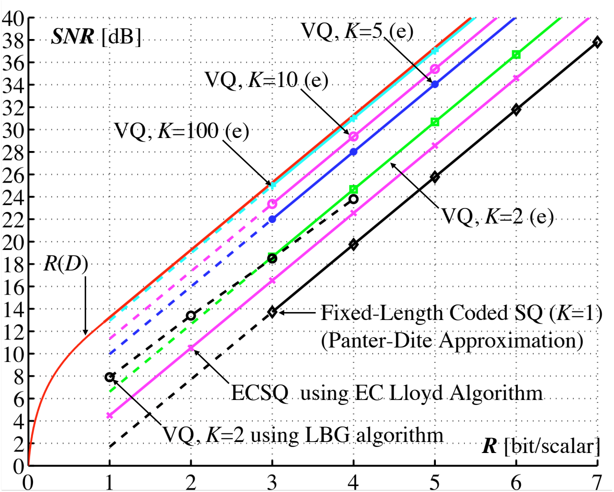
Summary on Memory Advantage

- Largest gain to be made if source contains statistical dependencies
- Exploiting the memory advantage is one of the most relevant aspects of source coding (shape advantage can be obtained using entropy coding)
- Remainder of source coding course will consider this issue



Vector Quantizer Advantage for a Gauss-Markov Source

- Gauss-Markov source with correlation factor $\rho = 0.9$
- Conjectured numbers are empirically verified for $K = 2$



Vector Quantization with Structural Constraints

- Vector quantizers can asymptotically achieve rate-distortion curve for $N \to \infty$
- Complexity requirements: Storage and computation
- Delay
- Impose structural constraints can reduce complexity
  - Tree-Structured VQ
  - Transform VQ
  - Multistage VQ
  - Shape-Gain VQ
  - Lattice Codebook VQ
  - Predictive VQ
- Predictive VQ can be seen as a generalization of very popular techniques: Motion compensation in video coding and various techniques in speech coding

Chapter Summary

Scalar quantization

- Lloyd quantizer: Minimum distortion for given number of representative levels
- Variable length coding: Additional gains by entropy-constrained quantization
- Optimal scalar quantizer for high rates: Uniform quantizer

Vector quantization

- Vector quantizers can achieve rate-distortion curve for $N \to \infty$
- Space filling gain: Only exploited by vector quantizers (1.53 dB for $N \to \infty$)
- Shape gain: Can also be exploited by entropy coding of quantization indices
- Memory gain: Can be exploited by predictive coding, transform coding, or entropy coding using joint or conditional probability mass functions

Vector quantization can achive rate-distortion bound. – Are we done?

$\Longrightarrow$ No! Complexity of vector quantizers is the issue

$\Longrightarrow$ **Design a coding system with optimum rate distortion performance, such that the delay, complexity, and storage requirements are met**

Exercise 15

Consider a symmetric scalar quantizer with 3 intervals and a quantizer input with a zero-mean Laplace pdf,

$$q(x) = \begin{cases} -b & : \quad x < -a \\ 0 & : \quad |x| \leq a \\ b & : \quad x > a \end{cases} \qquad\qquad f(x) = \frac{1}{2m} e^{-\frac{|x|}{m}}$$

(a) Derive the optimal reconstruction value b as a function of the decision threshold a for MSE distortion.

Express the resulting distortion as function of a and the variance $\sigma^2 = 2m^2$.

(b) Determine the decision threshold a in a way that a Lloyd quantizer for MSE distortion is obtained.

Determine the distortion and rate for the Lloyd quantizer by assuming fixed-length coding ($R = \log_2 N$) and compare the obtained R-D point with the Shannon lower bound.

(c) Can the derived optimal quantizer for fixed-length coding be improved by adding entropy coding (without changing the decision thresholds and reconstruction levels)?

Exercise 16

Given is a Centroidal quantizer (not necessarily a Lloyd quantizer) for MSE distortion and a source X. The quantizer has 5 reconstruction levels $\{-3, -1, 0, 1, 3\}$ which are chosen with probabilities $\{0.05, 0.1, 0.4, 0.3, 0.15\}$ and achieves an MSE of 1.05.

(a) Determine the mean μ and variance σ^2 of the source X.

(b) With $q(X)$ being the quantizer output and $e(X) = X - q(X)$ being the quantization error, determine the correlations $E\{X\,q(X)\}$, $E\{X\,e(X)\}$, and $E\{q(X)\,e(X)\}$.

Exercise 17

Consider a discrete Markov process $\mathbf{X} = \{X_n\}$ with the symbol alphabet $\mathcal{A}_X = \{0, 2, 4, 6\}$ and the conditional pmf

$$p_{X_n|X_{n-1}}(x_n|x_{n-1}) = \begin{cases} a & : \quad x_n = x_{n-1} \\ \frac{1}{3}(1-a) & : \quad x_n \neq x_{n-1} \end{cases},$$

for $x_n, x_{n-1} \in \mathcal{A}_X$. The parameter a, with $0 < a < 1$, is a variable that specifies the probability that the current symbol is equal to the previous symbol. For $a = 1/4$, our source $\mathbf{X}$ would be i.i.d.

Given is a quantizer of size 2 with the reconstruction levels $s'_0 = 1$ and $s'_1 = 5$ and the decision threshold $u_1 = 3$.

(a) Assume optimal entropy coding using the marginal probabilities of the quantization indices and determine the rate-distortion point of the quantizer.

(b) Can the overall quantizer performance be improved by applying conditional entropy coding (e.g., using arithmetic coding with conditional probabilities)? How does it depend on the parameter a?

Exercise 18

Calculate the gain of optimal 2-dimensional vector quantization relative to optimal scalar quantization for high rates on the example of a uniform pdf.

Hint:
For high rates, border effects can be neglected. It can be assumed that the signal space for which the pdf is non-zero is completely filled with regular quantization cells.

Exercise 19

Consider scalar quantization of a Laplacian source at high rates:

$$f(x) = \frac{\lambda}{2} \cdot e^{-\lambda\,|x|} \qquad \text{with} \qquad \sigma_S^2 = \frac{2}{\lambda^2}$$

In a given system, the used quantizer is a Lloyd quantizer with fixed-length entropy coding (the number of quantization intervals represents a power of 2). How many bits per sample can be saved if the quantizer is replaced by an entropy-constrained quantizer with optimal entropy coding?

Note: The operation points of the quantizers can be accurately described by high rate approximations.