EE-597 Notes – Quantization

Phil Schniter

June 11, 2004

1 Quantization

- Given a continuous-time and continuous-amplitude signal $x(t)$, processing and storage by modern digital hardware requires discretization in both time and amplitude, as accomplished by an analog-to-digital converter (ADC).
- We will typically work with discrete-time quantities $x(n)$ (indexed by "time" variable n) which we assume were sampled ideally and without aliasing.
- Here we discuss various ways of discretizing the amplitude of $x(n)$ so that it may be represented by a finite set of numbers. This is generally a lossy operation, and so we analyze the performance of various quantization schemes and design quantizers that are optimal under certain assumptions. A good reference for much of this material is [1].

1.1 Memoryless Scalar Quantization

- *Memoryless scalar quantization* of continuous-amplitude variable x is the mapping of x to output y_k when x lies within interval

$$\mathcal{X}_k := \{x_k < x \leq x_{k+1}\},\ k = 1, 2, \ldots, L.$$

  The x_k are called *decision thresholds*, and the number of quantization levels is L. The quantization operation is written $y = Q(x)$.

- When $0 \in \{y_1, \ldots, y_L\}$, quantizer is called *midtread*, else *midrise*.
- *Quantization error* defined $q := x - Q(x)$
- If x is a r.v. with pdf $p_x(\cdot)$ and likewise for q, then quantization error variance is

$$\begin{aligned} \sigma_q^2 \;=\; \mathrm{E}\{q^2\} &= \int_{-\infty}^{\infty} q^2 p_q(q) dq \qquad (1) \\ &= \int_{-\infty}^{\infty} \left(x - Q(x)\right)^2 p_x(x) dx \\ &= \sum_{k=1}^{L} \int_{x_k}^{x_{k+1}} (x - y_k)^2 p_x(x) dx \qquad (2) \end{aligned}$$

- A special quantizer is the *uniform quantizer*:

$$\begin{aligned} y_{k+1} - y_k &= \Delta, \quad \text{for } k = 1, 2, \ldots, L-1, \\ x_{k+1} - x_k &= \Delta, \quad \text{for finite } x_k, x_{k+1}, \\ -x_1 \;=\; x_{L+1} &= \infty. \end{aligned}$$

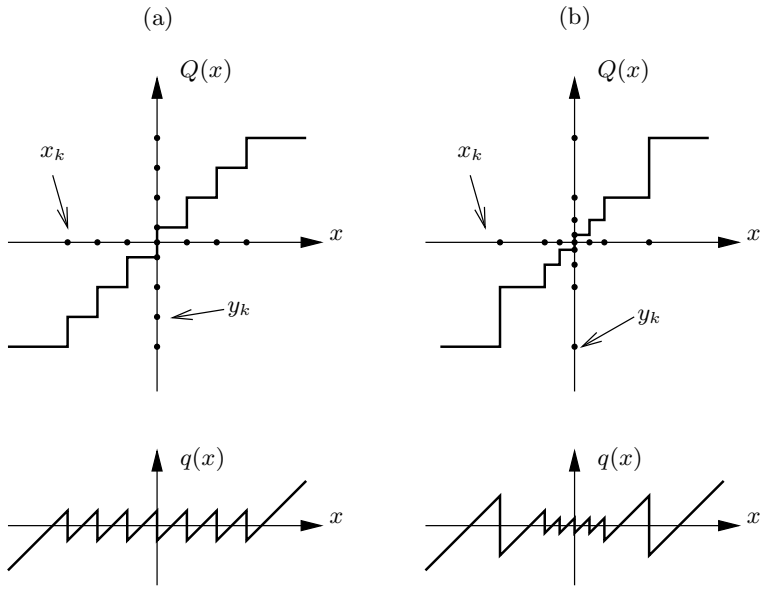


Figure 1: (a) Uniform and (b) non-uniform quantization $Q(x)$ and quantization error $q(x)$.

- *Uniform Quantizer Performance for large L:* For bounded input $x \in (-x_{\max}, x_{\max})$, uniform quantization with $x_2 = -x_{\max} + \Delta$ and $x_L = x_{\max} - \Delta$, and with $y_1 = x_2 - \Delta/2$ and $y_k = x_k + \Delta/2$ (for $k > 1$), the quantization error is well approximated by a uniform distribution for large L:

$$p_q(q) \;=\; \begin{cases} 1/\Delta & |q| \le \Delta/2, \\ 0 & \text{else.} \end{cases}$$

  Why?

  - As $L \to \infty$, $p_x(x)$ is constant over $\mathcal{X}_k$ for any k. Since $q = x - y_k\big|_{x \in \mathcal{X}_k}$, it follows that $p_q(q|x \in \mathcal{X}_k)$ will have uniform distribution for any k.
  - With $x \in (-x_{\max}, x_{\max})$ and with x_k and y_k as specified, $q \in (-\Delta/2, \Delta/2]$ for all x (see Fig. 2). Hence, for any k,

$$p_q(q|x \in \mathcal{X}_k) = \begin{cases} 1/\Delta & q \in (-\Delta/2, \Delta/2], \\ 0 & \text{else.} \end{cases}$$

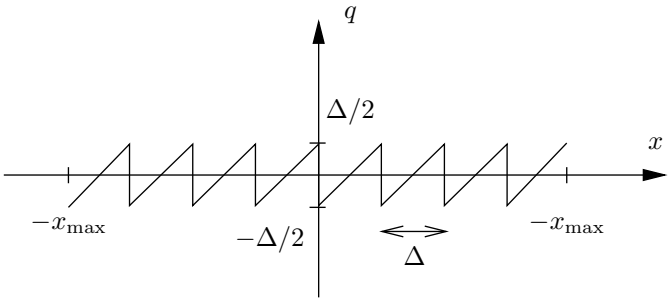


Figure 2: Quantization error for bounded input and midpoint y_k.

In this case, from (1),

$$\sigma_q^2 \;=\; \int_{-\Delta/2}^{\Delta/2} q^2 \frac{1}{\Delta}\,dq \;=\; \frac{1}{\Delta}\left[\frac{q^3}{3}\right]_{-\Delta/2}^{\Delta/2} \;=\; \frac{1}{\Delta}\left(\frac{\Delta^3}{3\cdot 8}+\frac{\Delta^3}{3\cdot 8}\right) \;=\; \boxed{\frac{\Delta^2}{12}}. \tag{3}$$

If we use R bits to represent each discrete output y and choose $L = 2^R$, then

$$\sigma_q^2 \;=\; \frac{\Delta^2}{12} \;=\; \frac{1}{12}\left(\frac{2x_{\max}}{L}\right)^2 \;=\; \frac{1}{3}x_{\max}^2\, 2^{-2R}$$

and

$$\text{SNR [dB]} \;=\; 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_q^2}\right) \;=\; 10\log_{10}\left(3\frac{\sigma_x^2}{x_{\max}^2}2^{2R}\right) \;=\; \boxed{6.02R - 10\log_{10}\left(3\frac{x_{\max}^2}{\sigma_x^2}\right)}.$$

Recall that the expression above is only valid for σ_x^2 small enough to ensure $x \in (-x_{\max}, x_{\max})$. For larger σ_x^2, the quantizer *overloads* and the SNR decreases rapidly.

Example 1.1 (SNR for Uniform Quantization of Uniformly-Distributed Input):
For uniformly distributed x, can show $x_{\max}/\sigma_x = \sqrt{3}$, so that $\text{SNR} = 6.02R$.

Example 1.2 (SNR for Uniform Quantization of Sinusoidal Input):
For a sinusoidal x, can show $x_{\max}/\sigma_x = \sqrt{2}$, so that $\text{SNR} = 6.02R + 1.76$. (Interesting since sine waves are often used as test signals).

Example 1.3 (SNR for Uniform Quantization of Gaussian Input):
Though not truly bounded, Gaussian x might be considered as approximately bounded if we choose $x_{\max} = 4\sigma_x$ and ignore residual clipping. In this case $\text{SNR} = 6.02R - 7.27$.

1.2 MSE-Optimal Memoryless Scalar Quantization

- Though uniform quantization is convenient for implementation and analysis, non-uniform quantization yeilds lower σ_q^2 when $p_x(\cdot)$ is non-uniformly distributed. By decreasing $|q(x)|$ for frequently occuring x (at the expense of increasing $|q(x)|$ for infrequently occuring x), the average error power can be reduced.

- <u>*Lloyd-Max Quantizer:*</u> MSE-optimal thresholds $\{x_k\}$ and outputs $\{y_k\}$ can be determined given an input distribution $p_x(\cdot)$, and the result is the *Lloyd-Max quantizer*. Necessary conditions on $\{x_k\}$ and $\{y_k\}$ are

$$\frac{\partial\sigma_q^2}{\partial x_k} = 0 \;\text{ for }\; k \in \{2,\ldots,L\} \quad\text{and}\quad \frac{\partial\sigma_q^2}{\partial y_k} = 0 \;\text{ for }\; k \in \{1,\ldots,L\}.$$

Using (2), $\partial/\partial b \int_a^b f(x)dx = f(b)$, $\partial/\partial a \int_a^b f(x)dx = -f(a)$, and above,

$$\frac{\partial\sigma_q^2}{\partial x_k} = (x_k - y_{k-1})^2 p_x(x_k) - (x_k - y_k)^2 p_x(x_k) = 0 \;\Rightarrow\; \boxed{\begin{array}{l} x_k^\star = \frac{y_k^\star + y_{k-1}^\star}{2},\; k \in \{2\ldots L\}, \\ x_1^\star = -\infty, \quad x_{L+1}^\star = \infty, \end{array}} \tag{4}$$

$$\frac{\partial\sigma_q^2}{\partial y_k} \;=\; 2\int_{x_k}^{x_{k+1}} (x - y_k)p_x(x)dx \;=\; 0 \;\Rightarrow\; \boxed{y_k^\star = \frac{\int_{x_k^\star}^{x_{k+1}^\star} x p_x(x)dx}{\int_{x_k^\star}^{x_{k+1}^\star} p_x(x)dx},\; k \in \{1\ldots L\}} \tag{5}$$

It can be shown that above are sufficient for global MMSE when $\partial^2 \log p_x(x)/\partial x^2 \le 0$, which holds for uniform, Gaussian, and Laplace pdfs, but not Gamma.

Note:

– optimum decision thresholds are halfway between neighboring output values,

– optimum output values are centroids of the pdf within the appropriate interval, i.e., are given by the conditional means

$$y_k^\star \;=\; \mathrm{E}\{x|x\in\mathcal{X}_k^\star\} \;=\; \int x p_x(x|x\in\mathcal{X}_k^\star)dx \;=\; \int x\frac{p_x(x,x\in\mathcal{X}_k^\star)}{\Pr(x\in\mathcal{X}_k^\star)}dx, \;=\; \frac{\int_{x_k^\star}^{x_{k+1}^\star} x p_x(x)dx}{\int_{x_k^\star}^{x_{k+1}^\star} p_x(x)dx}.$$

Iterative Procedure to Find $\{x_k^\star\}$ and $\{y_k^\star\}$:

1. Choose $\hat{y}_1$.
2. For $k=1,\ldots,L-1$,
   given $\hat{y}_k$ and $\hat{x}_k$, solve (5) for $\hat{x}_{k+1}$,
   given $\hat{y}_k$ and $\hat{x}_{k+1}$, solve (4) for $\hat{y}_{k+1}$.
   end;
3. Compare $\hat{y}_L$ to y_L calculated from (5) based on $\hat{x}_L$ and $x_{L+1}=\infty$. Adjust $\hat{y}_1$ accordingly, and go to step 1.

- <u>*Lloyd-Max Performance for large L:*</u> As with the uniform quantizer, can analyze quantization error performance for large L. Here, we assume that

  – the pdf $p_x(x)$ is constant over $x\in\mathcal{X}_k$ for $k\in\{1,\ldots,L\}$,

  – the input is bounded, i.e., $x\in(-x_{\max},x_{\max})$ for some (potentially large) $x_{\max}$.

  So with assumption

$$p_x(x)=p_x(y_k) \;\text{ for }\; x,y_k\in\mathcal{X}_k$$

  and definition

$$\Delta_k \;:=\; x_{k+1}-x_k,$$

  we can write

$$P_k \;:=\; \Pr\{x\in\mathcal{X}_k\} \;=\; p_x(y_k)\Delta_k \quad \Big(\text{where we require}\sum P_k=1\Big)$$

  and thus, from (2), σ_q^2 becomes

$$\sigma_q^2 \;=\; \sum_{k=1}^{L}\frac{P_k}{\Delta_k}\int_{x_k}^{x_{k+1}}(x-y_k)^2dx. \tag{6}$$

  For MSE-optimal $\{y_k\}$, know

$$0 \;=\; \frac{\partial\sigma_q^2}{\partial y_k} \;=\; 2\frac{P_k}{\Delta_k}\int_{x_k}^{x_{k+1}}(x-y_k)dx \quad\Rightarrow\quad \boxed{y_k^\star \;=\; \frac{x_k+x_{k+1}}{2}},$$

  which is expected since the centroid of a flat pdf over $\mathcal{X}_k$ is simply the midpoint of $\mathcal{X}_k$. Plugging $y_k^\star$ into expression (6),

$$\begin{aligned}\sigma_q^2 &= \sum_{k=1}^{L}\frac{P_k}{3\Delta_k}\left[(x-x_k/2-x_{k+1}/2)^3\right]_{x_k}^{x_{k+1}}\\ &= \sum_{k=1}^{L}\frac{P_k}{3\Delta_k}\left[(x_{k+1}/2-x_k/2)^3-(x_k/2-x_{k+1}/2)^3\right]\\ &= \sum_{k=1}^{L}\frac{P_k}{3\Delta_k}\cdot 2\left(\frac{\Delta_k}{2}\right)^3 \;=\; \frac{1}{12}\sum_{k=1}^{L}P_k\Delta_k^2.\end{aligned} \tag{7}$$

Note that for uniform quantization ($\Delta_k = \Delta$), the expression above reduces to the one derived earlier.

Now we minimize σ_q^2 w.r.t. $\{\Delta_k\}$. The trick here is to define

$$\alpha_k \;:=\; \sqrt[3]{p_x(y_k^\star)}\Delta_k \quad \text{so that} \quad \sigma_q^2 \;=\; \frac{1}{12}\sum_{k=1}^{L} p_x(y_k^\star)\Delta_k^3 \;=\; \frac{1}{12}\sum_{k=1}^{L}\alpha_k^3.$$

For $p_x(x)$ constant over $\mathcal{X}_k$ and $y_k \in \mathcal{X}_k$,

$$\sum_{k=1}^{L}\alpha_k \;=\; \sum_{k=1}^{L}\sqrt[3]{p_x(y_k^\star)}\Delta_k\Big|_{y_k^\star=\frac{x_k+x_{k+1}}{2}} \;=\; \int_{-x_{\max}}^{x_{\max}}\sqrt[3]{p_x(x)}dx \;=\; C_x \;\text{ (a known constant)},$$

we have the following constrained optimization problem:

$$\min_{\{\alpha_k\}}\sum_k \alpha_k^3 \quad \text{s.t.} \quad \sum_k \alpha_k = C_x.$$

This may be solved using Lagrange multipliers.

Aside 1.1 (Optimization via Lagrange Multipliers):
Consider the problem of minimizing N-dimensional real-valued cost function $J(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \ldots, x_N)^t$, subject to $M < N$ real-valued equality constraints $f_m(\mathbf{x}) = a_m$, $m = 1, \ldots, M$. This may be converted into an unconstrained optimization of dimension $N+M$ by introducing additional variables $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)^t$ known as Lagrange multipliers. The uncontrained cost function is

$$J_u(\mathbf{x}, \boldsymbol{\lambda}) \;=\; J(\mathbf{x}) + \sum_m \lambda_m\big(f_m(\mathbf{x}) - a_m\big),$$

and necessary conditions for its minimization are

$$\frac{\partial}{\partial \mathbf{x}}J_u(\mathbf{x},\boldsymbol{\lambda}) = \mathbf{0} \quad\Leftrightarrow\quad \frac{\partial}{\partial \mathbf{x}}J(\mathbf{x}) + \sum_m \lambda_m \frac{\partial}{\partial \mathbf{x}} f_m(\mathbf{x}) = \mathbf{0} \tag{8}$$

$$\frac{\partial}{\partial \boldsymbol{\lambda}}J_u(\mathbf{x},\boldsymbol{\lambda}) = \mathbf{0} \quad\Leftrightarrow\quad f_m(\mathbf{x}) = a_m \;\text{ for }\; m = 1, \ldots, M. \tag{9}$$

The typical procedure used to solve for optimal $\mathbf{x}$ is the following:

1. Equations for x_n, $n = 1, \ldots, N$, in terms of $\{\lambda_m\}$ are obtained from (8).
2. These N equations are used in (9) to solve for the M optimal λ_m.
3. The optimal $\{\lambda_m\}$ are plugged back into the N equations for x_n, yielding optimal $\{x_n\}$.

Necessary conditions are

$$\forall \ell,\; \frac{\partial}{\partial \alpha_\ell}\left(\sum_k \alpha_k^3 + \lambda\Big(\sum_k \alpha_k - C_x\Big)\right) = 0 \quad\Rightarrow\quad \lambda = -3\alpha_\ell^2 \quad\Rightarrow\quad \alpha_\ell = \sqrt{-\lambda/3}$$

$$\frac{\partial}{\partial \lambda}\left(\sum_k \alpha_k^3 + \lambda\Big(\sum_k \alpha_k - C_x\Big)\right) = 0 \quad\Rightarrow\quad \sum_k \alpha_k = C_x,$$

which can be combined to solve for λ:

$$\sum_{k=1}^{L}\sqrt{-\frac{\lambda}{3}} = C_x \quad\Rightarrow\quad \lambda = -3\left(\frac{C_x}{L}\right)^2.$$

Plugging λ back into the expression for α_ℓ, we find

$$\alpha_\ell^\star \;=\; C_x/L, \quad \forall\, \ell.$$

Using the definition of α_ℓ, the optimal decision spacing is

$$\Delta_k^\star \;=\; \frac{C_x}{L\sqrt[3]{p_x(y_k^\star)}} \;=\; \boxed{\frac{\int_{-x_{\max}}^{x_{\max}}\sqrt[3]{p_x(x)}dx}{L\sqrt[3]{p_x(y_k^\star)}}},$$

and the minimum quantization error variance is

$$\sigma_q^2\big|_{\min} \;=\; \frac{1}{12}\sum_k p_x(y_k^\star)\Delta_k^{\star\,3} \;=\; \frac{1}{12}\sum_k p_x(y_k^\star)\frac{\left(\int_{-x_{\max}}^{x_{\max}}\sqrt[3]{p_x(x)}dx\right)^3}{L^3 p_x(y_k^\star)}$$

$$= \;\boxed{\frac{1}{12L^2}\left(\int_{-x_{\max}}^{x_{\max}}\sqrt[3]{p_x(x)}dx\right)^3}.$$

An interesting observation is that $\alpha_\ell^{\star\,3}$, the ℓ^{th} interval's optimal contribution to σ_q^2, is constant over ℓ.

1.3 Entropy-Coding

- *Binary Scalar Encoding:* Previously we have focused on the memoryless scalar quantizer $y = Q(x)$, where y takes a value from a set of L reconstruction levels. By coding each quantizer output in binary format, we transmit (store) the information at a rate (cost) of

  $$R = \lceil \log_2 L \rceil \;\text{ bits/sample.}$$

  If, for example, $L = 8$, then we transmit at 3 bits/sample.

  Say we can tolerate a bit more quantization error, e.g., as results from $L = 5$. We hope that this reduction in fidelity reduces our transmission requirements, but with this simple binary encoding scheme we still require $R = 3$ bits/sample!

- *Idea—Block Coding:* Let's assign a symbol to each block of 3 consecutive quantizer outputs. We need a symbol alphabet of size $\geq 5^3 = 125$, which is adequately represented by a 7-bit word ($2^7 = 128$). Transmitting these words requires only $7/3 = 2.33$ bits/sample!

- *Idea—Variable Length Coding:* Assume some of the quantizer outputs occur more frequently than others. Could we come up with an alphabet consisting of short words for representing frequent outputs and longer words for infrequent outputs that would have a lower *average* transmission rate?

Example 1.4 (Variable Length Coding):

Consider the quantizer with $L = 4$ and output probabilities indicated on the right. Straightforward 2-bit encoding requires average bit rate of 2 bits/sample, while the variable length code on the right gives average $R = \sum_k P_k n_k = 0.6{\cdot}1 + 0.25{\cdot}2 + 0.1{\cdot}3 + 0.05{\cdot}3 = 1.55$ bits/sample.

output	P_k	code
y_1	0.60	0
y_2	0.25	01
y_3	0.10	011
y_4	0.05	111

- *(Just enough information about) Entropy:*

  Q: Given an arbitrarily complex coding scheme, what is the minimum bits/sample required to transmit (store) the sequence $\{y(n)\}$?

  A: When random process $\{y(n)\}$ is i.i.d., the minimum average bit rate is

  $$R_{\min} \;=\; H_y + \epsilon,$$

where H_y is the *entropy* of random variable $y(n)$ in bits:

$$H_y \;=\; -\sum_{k=1}^{L} P_k \log_2 P_k,$$

and ϵ is an arbitrarily small positive constant [2], [3].

Notes:

- Entropy obeys $0 \le H_y \le \log_2 L$. The left inequality occurs when $P_k = 1$ for some k, while the right inequality occurs when $P_k = 1/L$ for every k.
- The term *entropy* refers to the average information of a single random variable, while the term *entropy rate* refers to a sequence of random variables, i.e., a random process.
- When $\{y(n)\}$ is not independent (the focus of later sections), a different expression for $R_{\min}$ applies.
- Though the minimum rate is well specified, the construction of a coding scheme which always achieves this rate is not.

Example 1.5 (Entropy of Variable Length Code):

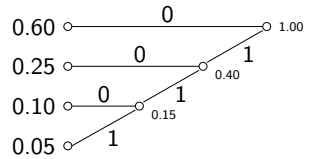
Recalling the setup of Example 1.4, we find that $H_y = -\big(0.6\log_2 0.6 + 0.25\log_2 0.25 + 0.1\log_2 0.1 + 0.05\log_2 0.05\big) = 1.49$ bits. Assuming i.i.d. $\{y(n)\}$, we have $R_{\min} = 1.49$ bits/sample. Compare to the variable length code on the right which gave $R = 1.55$ bits/sample.

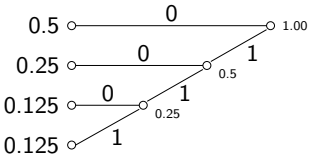
output	P_k	code
y_1	0.60	0
y_2	0.25	01
y_3	0.10	011
y_4	0.05	111

- <u>*Huffman Encoding:*</u> Given quantizer outputs y_k or fixed-length blocks of outputs $(y_j\, y_k\, y_\ell)$, the *Huffman* procedure constructs variable length codes that are optimal in certain respects [3]. For example, when the probabilities of $\{P_k\}$ are powers of 1/2 (and $\{y(n)\}$ is i.i.d.), the entropy rate of a Huffman encoded output attains $R_{\min}$.

Aside 1.2 (Huffman Procedure (Binary Case)):

1. Arrange ouput probabilities P_k in decreasing order and consider them as leaf nodes of a tree.
2. While there exists more than one node:
   - Merge the two nodes with smallest probability to form a new node whose probability equals the sum of the two merged nodes.
   - Arbitrarily assign 1 and 0 to the two branches of the merging pair.
3. The code assigned to each output is obtained by reading the branch bits sequentially from root note to leaf node.





Example 1.6 (Huffman Encoder Attaining $R_{\min}$):

In Aside 1.2, a Huffman code was constructed for the output probabilities listed on the right. Here $H_y = -\big(0.5\log_2 0.5 + 0.25\log_2 0.25 + 2\cdot 0.125\log_2 0.125\big) = 1.75$ bits, so that $R_{\min} = 1.75$ bits/sample (with the i.i.d. assumption). Since the average bit rate for the Huffman code is also $R = 0.5{\cdot}1 + 0.25{\cdot}2 + 0.125{\cdot}3 + 0.125{\cdot}3 = 1.75$ bits/sample, Huffman encoding attains $R_{\min}$ for this output distribution.

output	P_k	code
y_1	0.5	0
y_2	0.25	01
y_3	0.125	011
y_4	0.125	111

1.4 Quantizer Design for Entropy Coded Sytems

- Say that we are designing a system with a memoryless quantizer followed by an entropy coder, and our goal is to minimize the average transmission rate for a given σ_q^2 (or vice versa). Is it optimal to cascade a σ_q^2-minimizing (Lloyd-Max) quantizer with a rate-minimizing code? In other words, what is the optimal memoryless quantizer if the quantized outputs are to be entropy coded?

- *A Compander Formulation:*
  To determine the optimal quantizer,
  1. consider a *companding* system: a memoryless nonlinearity $c(x)$ followed by uniform quantizer,
  2. find $c(x)$ minimizing entropy H_y for a fixed error variance σ_q^2.

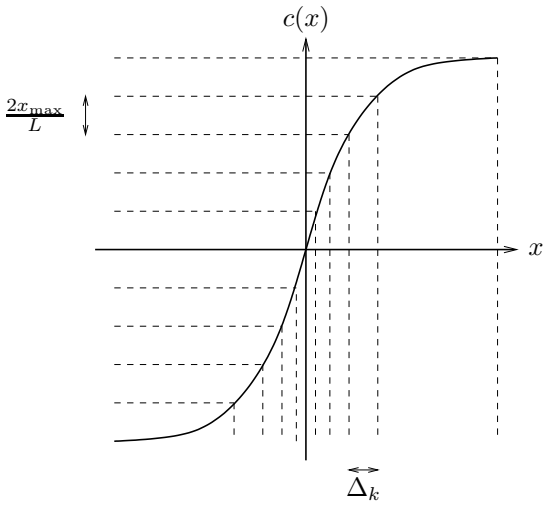


Figure 3: Compander curve: nonuniform input regions mapped to uniform output regions (for subsequent uniform quantization).

- First we must express σ_q^2 and H_y in terms of $c(x)$. Fig. 3 suggests that, for large L, the slope $c'(x) := dc(x)/dx$ obeys

$$c'(x)\big|_{x\in\mathcal{X}_k} = \frac{2x_{\max}/L}{\Delta_k},$$

  so that we may write

$$\Delta_k = \frac{2x_{\max}}{Lc'(x)}\bigg|_{x\in\mathcal{X}_k}.$$

  Assuming large L, the σ_q^2-approximation (7) can be transformed as follows.

$$\begin{aligned}
\sigma_q^2 &= \frac{1}{12}\sum_{k=1}^{L} P_k\Delta_k^2 \\
&= \frac{x_{\max}^2}{3L^2}\sum_{k=1}^{L}\frac{P_k}{c'(x)^2}\bigg|_{x\in\mathcal{X}_k} \\
&= \frac{x_{\max}^2}{3L^2}\sum_{k=1}^{L}\frac{p_x(x)}{c'(x)^2}\Delta_k\bigg|_{x\in\mathcal{X}_k} \qquad \text{since} \quad P_k = p_x(x)\big|_{x\in\mathcal{X}_k}\Delta_k \quad \text{for large } L \\
&= \frac{x_{\max}^2}{3L^2}\int_{-x_{\max}}^{x_{\max}}\frac{p_x(x)}{c'(x)^2}dx.
\end{aligned}$$

Similarly,

$$\begin{aligned}
H_y &= -\sum_{k=1}^{L} P_k \log_2 P_k \\
&= -\sum_{k=1}^{L} p_x(x)\Delta_k \log_2\big(p_x(x)\Delta_k\big)\big|_{x\in\mathcal{X}_k} \\
&= -\sum_{k=1}^{L} p_x(x)\Delta_k \log_2 p_x(x)\big|_{x\in\mathcal{X}_k} - \sum_{k=1}^{L} p_x(x)\Delta_k \log_2 \Delta_k\big|_{x\in\mathcal{X}_k} \\
&= \underbrace{-\int_{-x_{\max}}^{x_{\max}} p_x(x)\log_2 p_x(x)dx}_{h_x\text{:``differential entropy''}} - \int_{-x_{\max}}^{x_{\max}} p_x(x)\log_2 \underbrace{\frac{2x_{\max}}{Lc'(x)}}_{\Delta_k} dx \\
&= h_x - \log_2\frac{2x_{\max}}{L}\underbrace{\int_{-x_{\max}}^{x_{\max}} p_x(x)dx}_{=1} + \int_{-x_{\max}}^{x_{\max}} p_x(x)\log_2 c'(x)dx \qquad (10) \\
&= \text{constant} + \int_{-x_{\max}}^{x_{\max}} p_x(x)\log_2 c'(x)dx
\end{aligned}$$

- *Entropy-Minimizing Quantizer:* Our goal is to choose $c(x)$ which minimizes the entropy rate H_y subject to fixed error variance σ_q^2. We employ a Lagrange technique again, minimizing the cost $\int_{-x_{\max}}^{x_{\max}} p_x(x)\log_2 c'(x)dx$ under the constraint that the quantity $\int_{-x_{\max}}^{x_{\max}} p_x(x)\big(c'(x)\big)^{-2}dx$ equals a constant C. This yields the unconstrained cost function

$$J_u\big(c'(x),\lambda\big) = \int_{-x_{\max}}^{x_{\max}} \underbrace{\Big[p_x(x)\log_2 c'(x) + \lambda\Big(p_x(x)\big(c'(x)\big)^{-2} - C\Big)\Big]}_{\phi(c'(x),\lambda)} dx, \qquad (11)$$

with scalar λ, and the unconstrained optimization problem becomes

$$\min_{c'(x),\lambda} J_u(c'(x),\lambda).$$

The following technique is common in variational calculus [4]. Say $a^\star(x)$ minimizes a (scalar) cost $J\big(a(x)\big)$. Then for *any* (well-behaved) variation $\eta(x)$ from this optimal $a^\star(x)$, we must have

$$\frac{\partial}{\partial\epsilon} J\big(a^\star(x)+\epsilon\eta(x)\big)\Big|_{\epsilon=0} = 0$$

where ϵ is a scalar. Applying this principle to our optimization problem, we search for $c'(x)$ such that

$$\forall\eta(x),\quad \frac{\partial}{\partial\epsilon} J_u\big(c'(x)+\epsilon\eta(x),\lambda\big)\Big|_{\epsilon=0} = 0.$$

From (11) we find (using $\log_2 a = \log_2 e\cdot\log_e a$)

$$\begin{aligned}
\frac{\partial J_u}{\partial\epsilon}\Big|_{\epsilon=0} &= \int_{-x_{\max}}^{x_{\max}} \frac{\partial}{\partial\epsilon}\phi\big(c'(x)+\epsilon\eta(x),\lambda\big)\Big|_{\epsilon=0} dx \\
&= \int_{-x_{\max}}^{x_{\max}} \frac{\partial}{\partial\epsilon}\Big[p_x(x)\log_2(e)\,\log_e\Big(c'(x)+\epsilon\eta(x)\Big) + \lambda\Big(p_x(x)\big(c'(x)+\epsilon\eta(x)\big)^{-2} - C\Big)\Big]\Big|_{\epsilon=0} dx \\
&= \int_{-x_{\max}}^{x_{\max}} \Big[\log_2(e)\,p_x(x)\Big(c'(x)+\epsilon\eta(x)\Big)^{-1}\eta(x) - 2\lambda p_x(x)\Big(c'(x)+\epsilon\eta(x)\Big)^{-3}\eta(x)\Big]\Big|_{\epsilon=0} dx \\
&= \int_{-x_{\max}}^{x_{\max}} p_x(x)\big(c'(x)\big)^{-1}\big[\log_2(e) - 2\lambda\big(c'(x)\big)^{-2}\big]\,\eta(x)\,dx
\end{aligned}$$

and to allow for any $\eta(x)$ we require

$$\log_2(e) - 2\lambda\big(c'(x)\big)^{-2} = 0 \quad \Leftrightarrow \quad c'(x) = \underbrace{\sqrt{\frac{2\lambda}{\log_2 e}}}_{\text{a constant!}}.$$

Applying the boundary conditions,

$$\left\{ \begin{array}{c} c(x_{\max}) = x_{\max} \\ c(-x_{\max}) = -x_{\max} \end{array} \right\} \quad \rightarrow \quad \boxed{c(x) = x}.$$

Thus, for large-L, *the quantizer that minimizes entropy rate H_y for a given quantization error variance σ_q^2 is the uniform quantizer.*

Plugging $c(x) = x$ into (10), the rightmost integral disappears and we have

$$H_y\big|_{\text{uniform}} \;=\; h_x - \log_2 \underbrace{\frac{2x_{\max}}{L}}_{\Delta},$$

and using the large-L uniform quantizer error variance approximation (3),

$$H_y\big|_{\text{uniform}} \;=\; h_x - \frac{1}{2}\log_2\big(12\sigma_q^2\big).$$

It is interesting to compare this result to the information-theoretic minimal average rate for transmission of a continuous-amplitude memoryless source x of differential entropy h_x at average distortion σ_q^2 [1, 2]:

$$R_{\min} \;=\; h_x - \frac{1}{2}\log_2\big(2\pi e\,\sigma_q^2\big).$$

Comparing the previous two equations, we find that (for a continous-amplitude memoryless source) uniform quantization prior to entropy coding requires

$$\frac{1}{2}\log_2\left(\frac{\pi e}{6}\right) \;\approx\; \boxed{0.255 \;\text{ bits/sample}}$$

more than the theoretically optimum transmission scheme, regardless of the distribution of x. Thus, *0.255 bits/sample (or $\sim$ 1.5 dB using the $6.02R$ relationship) is the price paid for memoryless quantization.*

1.5 Adaptive Quantization

- Previously have considered the case of stationary source processes, though in reality the source signal may be highly non-stationary. For example, the variance, pdf, and/or mean may vary significantly with time.

- Here we concentrate on the problem of adapting uniform quantizer stepsize Δ to a signal with unknown variance. This is accomplished by estimating the input variance $\hat{\sigma}_x(n)$ and setting the quantizer stepsize appropriately:

$$\Delta(n) = \frac{2\phi_x\,\hat{\sigma}_x(n)}{L}.$$

Here ϕ_x is a constant that depends on the distribution of the input signal x whose function is to prevent input values greater than $\sigma_x(n)$ from being clipped by the quantizer (see Fig. 4); comparing to non-adaptive step size relation $\Delta = 2x_{\max}/L$, we see that $\phi_x\hat{\sigma}_x(n) \sim x_{\max}$.

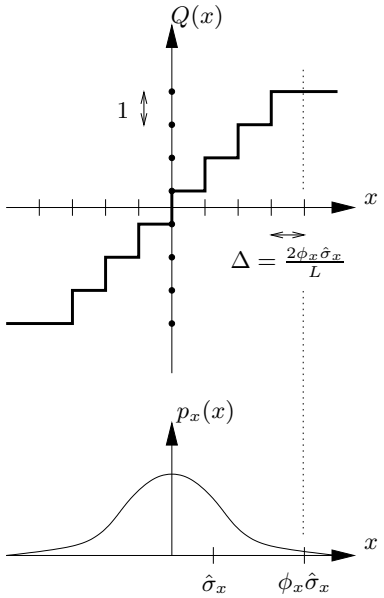


Figure 4: Adaptive quantization stepsize $\Delta(n) = 2\phi_x\hat{\sigma}_x/L$

- As long as the reconstruction levels $\{y_k\}$ are the same at encoder and decoder, the actual values chosen for quantizer design are arbitrary. Assuming integer values as in Fig. 4, the quantization rule becomes

$$y(n) \;=\; \begin{cases} \left\lceil \frac{x(n)}{\Delta(n)} \right\rceil - \frac{1}{2} & \text{midrise,} \\ \left\lceil \frac{x(n)}{\Delta(n)} - \frac{1}{2} \right\rceil & \text{midtread.} \end{cases} \tag{12}$$

- *AQF and AQB:* Fig. 5 shows two structures for stepsize adaptation: (a) *adaptive quantization with forward estimation* (AQF) and (b) *adaptive quantization with backward estimation* (AQB). The advantage of AQF is that variance estimation may be accomplished more accurately, as it is operates directly on the source as opposed to a quantized (noisy) version of the source. The advantage of AQB is that the variance estimates do not need to be transmitted as side information for decoding. In fact, practical AQF encoders transmit variance estimates only occasionally, e.g., once per block.

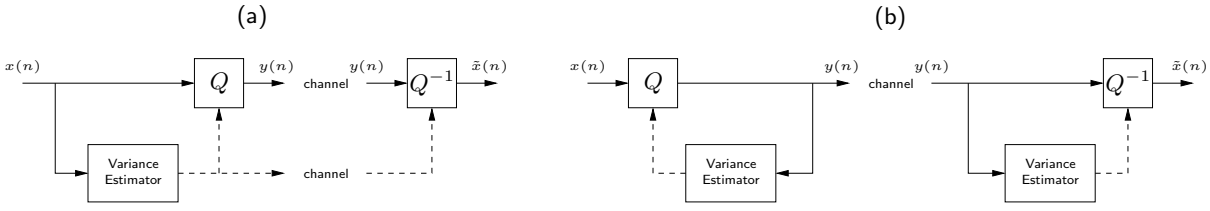


Figure 5: (a) AQF and (b) AQB.

- *Block Variance Estimation:* When operating on finite blocks of data, the structures in Fig. 5

perform variance estimation as follows:

$$\text{Block AQF:} \qquad \hat{\sigma}_x^2(n) \;=\; \frac{1}{N}\sum_{i=1}^{N} x^2(n-i)$$

$$\text{Block AQB:} \qquad \hat{\sigma}_x^2(n) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big(y(n-i)\cdot\Delta(n-i)\Big)^2$$

N is termed the *learning period* and its choice may significantly impact quantizer SNR performance: choosing N too large prevents the quantizer from adapting to the local statistics of the input, while choosing N too small results in overly noisy AQB variance estimates and excessive AQF side information. Fig. 6 demonstrates these two schemes for two choices of N.

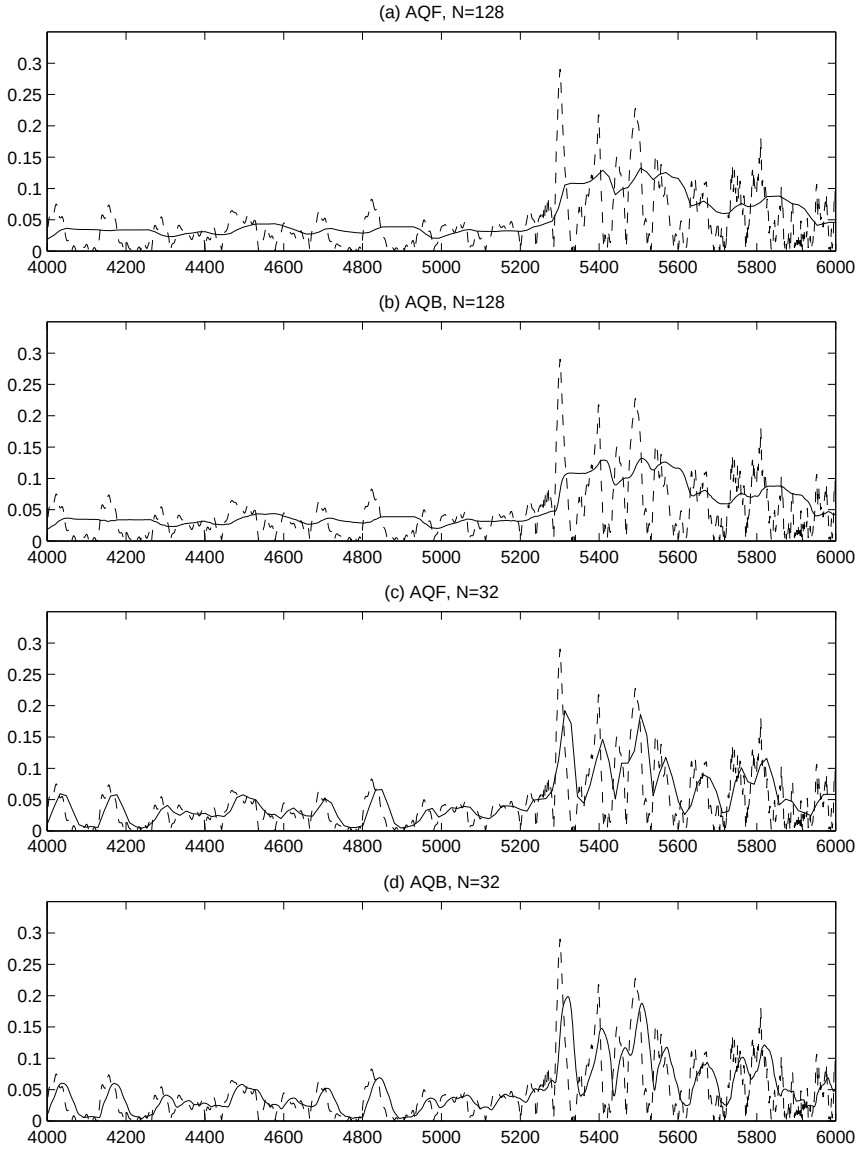


Figure 6: Block AQF and AQB estimates of $\sigma_x(n)$ superimposed on $|x(n)|$ for $N = 128, 32$. SNR acheived: (a) 22.6 dB, (b) 28.8 dB, (c) 21.2 dB, and (d) 28.8 dB.

- *Recursive Variance Estimation:* The recursive method of estimating variance is as follows

$$\begin{aligned}
\text{Recursive AQF:} \quad \hat{\sigma}_x^2(n) &= \alpha\hat{\sigma}_x^2(n-1) + (1-\alpha)x^2(n-1) \\
\text{Recursive AQB:} \quad \hat{\sigma}_x^2(n) &= \alpha\hat{\sigma}_x^2(n-1) + (1-\alpha)\Big(y(n-1)\cdot\Delta(n-1)\Big)^2.
\end{aligned}$$

where α is a *forgetting factor* in the range $0 < \alpha < 1$ and typically near to 1.

This leads to an exponential data window, as can be seen below. Plugging the expression for $\hat{\sigma}_x^2(n-1)$ into that for $\hat{\sigma}_x^2(n)$,

$$\begin{aligned}
\hat{\sigma}_x^2(n) &= \alpha\Big(\alpha\hat{\sigma}_x^2(n-2) + (1-\alpha)x^2(n-2)\Big) + (1-\alpha)x^2(n-1) \\
&= \alpha^2\hat{\sigma}_x^2(n-2) + (1-\alpha)\Big(x^2(n-1) + \alpha x^2(n-2)\Big).
\end{aligned}$$

Then plugging $\hat{\sigma}_x^2(n-2)$ into the above,

$$\begin{aligned}
\hat{\sigma}_x^2(n) &= \alpha^2\Big(\alpha\hat{\sigma}_x^2(n-3) + (1-\alpha)x^2(n-3)\Big) + (1-\alpha)\Big(x^2(n-1) + \alpha x^2(n-2)\Big) \\
&= \alpha^3\hat{\sigma}_x^2(n-3) + (1-\alpha)\Big(x^2(n-1) + \alpha x^2(n-2) + \alpha^2 x^2(n-3)\Big).
\end{aligned}$$

Continuing this process N times, we arrive at

$$\hat{\sigma}_x^2(n) = (1-\alpha)\sum_{i=1}^{N}\alpha^{i-1}x^2(n-i) + \alpha^N\hat{\sigma}_x^2(n-N).$$

Taking the limit as $N \to \infty$, $\alpha < 1$ ensures that

$$\boxed{\hat{\sigma}_x^2(n) = (1-\alpha)\sum_{i=1}^{\infty}\alpha^{i-1}x^2(n-i).}$$

References

[1] N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Englewood Cliffs, NJ: Prentice-Hall, 1984.

[2] T. Berger, *Rate Distortion Theory*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[3] T.A. Cover and J.A. Thomas, *Elements of Information Theory*, New York, NY: Wiley, 1991.

[4] A.P. Sage and C.C. White, III, *Optimum Systems Control, 2nd Ed.*, Englewood Cliffs, NJ: Prentice-Hall, 1977.

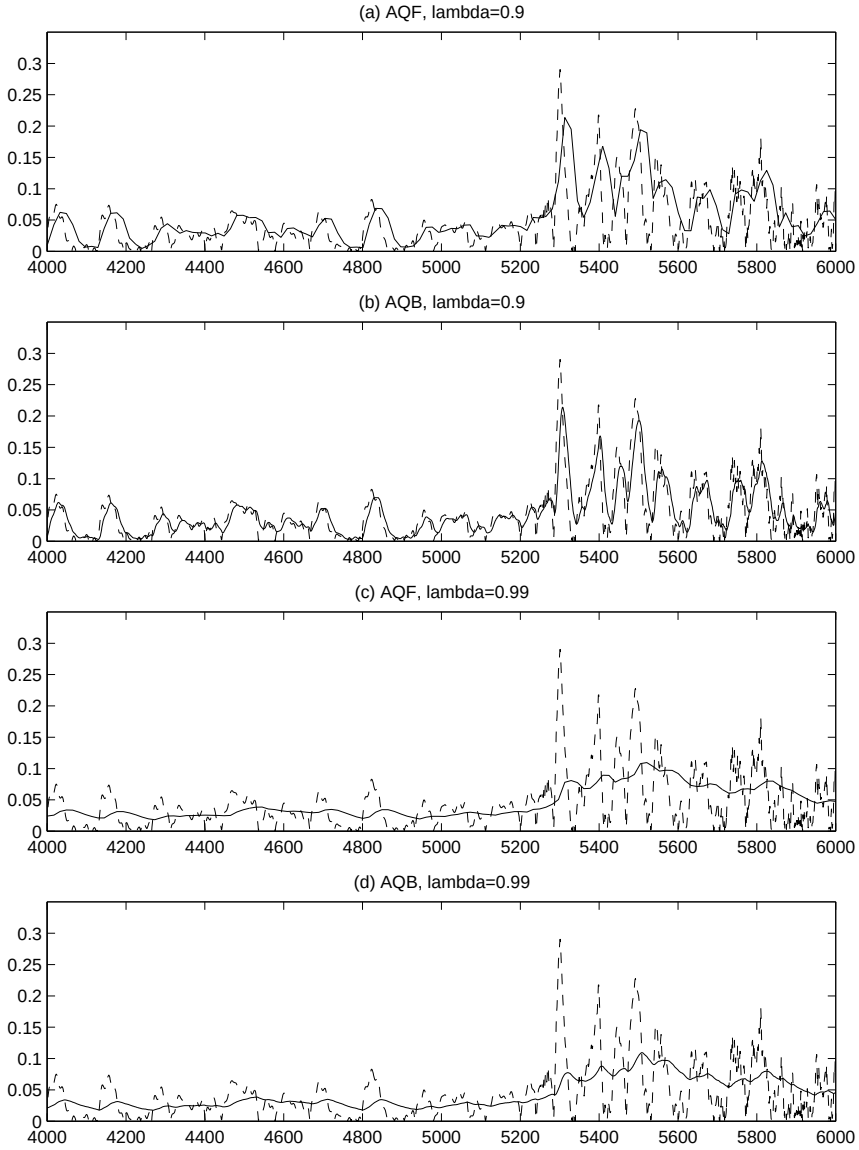


Figure 7: Exponential AQF and AQB estimates of $\sigma_x(n)$ superimposed on $|x(n)|$ for $\lambda = 0.9, 0.99$. (a) 20.5 dB, (b) 28.0 dB, (c) 22.2 dB, (d) 24.1 dB.