ENEE 420
FALL 2012
COMMUNICATIONS SYSTEMS

QUANTIZATION

Throughout, let X stand for a scalar rv taking values in the interval $I := (A, B]$ for some finite scalars $A < B$. We denote by F its probability distribution function, so that

$$\mathbb{P}[X \leq x] = F(x), \quad x \in \mathbb{R}.$$

We shall assume that F admits a probability density function f, so that

$$F(x) = \begin{cases} 0 & \text{if } x \leq A \\ \int_A^x f(t)dt & \text{if } A < x \leq B \\ 1 & \text{if } B < x. \end{cases}$$

Quantizers

A *quantizer* Q with M levels for the interval $(A, B]$ (or interchangeably, for any rv X distributed on the interval $(A, B]$) is characterized by a collection of M *contiguous* sub-intervals or *cells* partioning $(A, B]$, say $I_1, \ldots, I_M$, and a collection of *representation levels* $q_1, \ldots, q_M$, one to represent each of the intervals. The partitioning constraints amounts to

$$I_m := (A_m, B_m], \quad m = 1, \ldots M$$

with the notation

$$\begin{cases} A_1 = A; \\ A_{m+1} = B_m, \ m = 1, \ldots, M-1; \\ B_M = B. \end{cases}$$

We also require

$$q_m \in I_m, \quad m = 1, \ldots, M.$$

Often we shall denote such a quantizer Q by

$$Q \equiv (I_1, \ldots, I_M; q_1, \ldots, q_M).$$

At times it will also be convenient to think of this quantizer as a *mapping* $Q : I \to I$ given by

$$Q(x) = q_m \quad \text{if } x \in I_m, \quad m = 1, \ldots, M.$$

Uniform quantizers

A quantizer is said to be *uniform* on the interval $(A, B]$ if its cells all have the *same* length and the representatives are *equidistant*. This uniquely determines the quantizer $Q^u = (I_1^u, \ldots, I_M^u; q_1^u, \ldots, q_M^u)$, hereafter referred to as the *uniform quantizer* for the interval $(A, B]$, with

$$B_1^u - A_1^u = B_2^u - A_2^u = \ldots = B_M^u - A_M^u$$

and

$$q_m^u = \frac{A_m^u + B_m^u}{2}, \quad m = 1, \ldots, M.$$

Indeed, each interval must have length $\frac{B-A}{M}$ so that for each $m = 1, \ldots, M$

$$A_m^u = A + (m-1) \cdot \frac{B-A}{M}$$

and

$$B_m^u = A + m \cdot \frac{B-A}{M},$$

whence

$$q_m^u = \frac{A^u m + B_m^u}{2} = A + \frac{2m-1}{2} \cdot \frac{B-A}{M}.$$

Measuring distortion

If X is the variable to be quantized, then the *quantization error* or *quantization noise* under the quantizer Q is given by

$$\varepsilon(Q; X) := Q(X) - X.$$

With the quantizer Q we associate the *distortion measure*

$$\Phi(Q;F) := \mathbb{E}\left[|\varepsilon(Q;X)|^2\right] \tag{1}$$

as a way to assess how well the quantized version $Q(X)$ of X approximates X.

We define the *signal-to-quantization-noise ratio* (SQNR) associated with the quantizer Q as the ratio

$$\mathrm{SQNR}(Q;X) := \frac{\mathbb{E}\left[X^2\right]}{\mathbb{E}\left[|\varepsilon(Q;X)|^2\right]}.$$

In selecting a quantizer for the rv X it should be intuitively clear that a large value for $\mathrm{SQNR}(Q;X)$ is desirable.

The quantization problem

Fix some positive integer $M \geq 2$. Given the rv X distributed over the interval I, we are interested in minimizing the distortion measure (1) over all possible quantizers with M levels for the interval I.

For any such quantizer $Q \equiv (I_1, \ldots, I_M; q_1, \ldots, q_M)$, we note that

$$\begin{aligned}
\Phi(Q;F) &= \mathbb{E}\left[|\varepsilon(Q;X)|^2\right] \\
&= \int_{\mathbb{R}} |\varepsilon(Q;x)|^2 f(x)dx \\
&= \int_A^B |Q(x) - x|^2 f(x)dx \\
&= \sum_{m=1}^M \int_{A_m}^{B_m} |Q(x) - x|^2 f(x)dx \\
&= \sum_{m=1}^M \int_{A_m}^{B_m} |q_m - x|^2 f(x)dx. \qquad (2)
\end{aligned}$$

Thus, with quantizer Q characterized by cells $I_1, \ldots, I_m$ and representation levels $q_1, \ldots, q_M$, we shall write

$$\Phi(Q;F) = \Phi_F(I_1, \ldots, I_M; q_1, \ldots, q_M)$$

with

$$\Phi_F(I_1, \ldots, I_M; q_1, \ldots, q_M) = \sum_{m=1}^M \int_{A_m}^{B_m} |q_m - x|^2 f(x)dx. \tag{3}$$

Given cells $I_1, \ldots, I_M$

We start with contiguous cells $I_1, \ldots, I_M$ partitioning I, and focus on the following minimization problem: Find the representation levels $q_1, \ldots, q_M$ which minimize

$$\Phi_F(I_1, \ldots, I_M; q_1, \ldots, q_M)$$

under the constraints

$$q_m \in I_m,\ m = 1, \ldots, M.$$

The expression (3) is *separable* in the variables $q_1, \ldots, q_M$ and the constraints on them. As a result, the original minimization problem can be solved by solving each of the following M sub-problems. Indeed,

$$\begin{aligned} & \min \left(\sum_{m=1}^{M} \int_{A_m}^{B_m} |q_m - x|^2 f(x) dx,\ q_m \in I_m,\ m = 1, \ldots, M \right) \\ = & \sum_{m=1}^{M} \min \left(\int_{A_m}^{B_m} |q_m - x|^2 f(x) dx,\ q_m \in I_m \right). \end{aligned} \tag{4}$$

With this in mind, fix $m = 1, \ldots, M$. We now seek to minimize

$$\int_{A_m}^{B_m} |q_m - x|^2 f(x) dx$$

under the constraint

$$q_m \in I_m.$$

The solution is straightforward: We note that

$$\begin{aligned} & \int_{A_m}^{B_m} |q_m - x|^2 f(x) dx \\ = & q_m^2 \int_{A_m}^{B_m} f(x) dx - 2 q_m \int_{A_m}^{B_m} x f(x) dx + \int_{A_m}^{B_m} x^2 f(x) dx. \end{aligned} \tag{5}$$

This quadratic form in the variable q_m is minimized at $q_m^\star$ given by

$$q_m^\star = \frac{\int_{A_m}^{B_m} x f(x) dx}{\int_{A_m}^{B_m} f(x) dx}.$$

This can be seen by a completion-of-square argument, or by taking the derivative with respect the variable q_m and setting it equal to zero: Thus,

$$\begin{aligned} & \frac{1}{2}\frac{d}{dq_m}\left(\int_{A_m}^{B_m} |q_m - x|^2 f(x)dx\right) \\ = & \ q_m \int_{A_m}^{B_m} f(x)dx - \int_{A_m}^{B_m} xf(x)dx \\ = & \ 0, \end{aligned} \tag{6}$$

and the value for $q_m^\star$ follows. It is easy to see that

$$A_m \leq q_m^\star \leq B_m$$

and the candidate solution $q_m^\star$ obtained by unconstrained minimization is an element of I_m, as required.

Thus, for each $m = 1, 2, \ldots, M$, given the interval I_m, we have

$$\min\left(\int_{A_m}^{B_m} |q_m - x|^2 f(x)dx,\ q_m \in I_m\right) = \int_{A_m}^{B_m} |q_m^\star - x|^2 f(x)dx.$$

where

$$q_m^\star == \frac{\int_{A_m}^{B_m} xf(x)dx}{\int_{A_m}^{B_m} f(x)dx}.$$

Given representation levels $q_1, \ldots, q_M$

This time we are given M distinct representation levels in I, say $A < q_1 < \ldots < q_M < B$, and we focus on the following minimization problem: Find the cells $I_1, \ldots, I_M$ which minimize

$$\Phi_F(I_1, \ldots, I_M; q_1, \ldots, q_M)$$

under the constraints

$$I_1 \cup \ldots \cup I_M = (A, B]$$

and

$$q_m \in I_m,\ m = 1, \ldots, M.$$

In contrast with the problem discussed earlier, this minimization problem is no more separable. However, a careful inspection of the expression (3) suggests that the intervals

$$I_m^\star := \left\{ x \in (A,B] : \ |x-q_m|^2 \leq |x-q_k|^2, \quad \begin{array}{c} k=1,\ldots,M \\ k \neq m \end{array} \right\}, \quad m=1,\ldots,M$$

constitute the solution.[1] Before giving a proof of this assertion, it is worth pointing out that for distinct ℓ and m, the inequality

$$|x-q_m|^2 \leq |x-q_\ell|^2 \tag{7}$$

occurs if and only if

$$(q_m - q_\ell)(2x - (q_\ell + q_m)) \geq 0.$$

Thus, if $q_m < q_\ell$ (resp. $q_\ell < q_m$), then (7) holds provided $x \leq \frac{q_m+q_\ell}{2}$. Similarly, if $q_\ell < q_m$, then (7) holds provided $x \geq \frac{q_m+q_\ell}{2}$. A moment of reflection shows that the sets $I_1^\star, \ldots, I_m^\star$ are indeed *intervals* of the form

$$I_m^\star = (A_m^\star, B_m^\star], \quad m=1,\ldots,M$$

with

$$A_1^\star = A, \quad A_m^\star = \frac{q_{m-1}+q_m}{2}, \quad m=2,\ldots,M.$$

It goes without saying that $B_m^\star = A_{m+1}^\star$ for each $m=1,\ldots,M-1$ and $B_M^\star = B$.

To establish the optimality of the intervals $I_1^\star, \ldots, I_M^\star$, we proceed as follows: Recall that for any function $g : I \to \mathbb{R}$, the linearity of the intergral operation gives

$$\int_I g(x)dx = \sum_{m=1}^M \int_{I_m} g(x)dx$$

for any partition $I_1, \ldots, I_m$ of the interval I. Now, for each $m=1,\ldots,M$, the definition of the interval $I_m^\star$ yields

$$|x-q_m|^2 = \min_{k=1,\ldots,M} |x-q_k|^2, \quad x \in I_m^\star.$$

[1] The boundary points are selected so as to create intervals which are open to the left and closed to the right.

Therefore,

$$\begin{aligned} & \Phi_F(I_1,\ldots,I_M;q_1,\ldots,q_M) - \Phi_F(I_1^\star,\ldots,I_M^\star;q_1,\ldots,q_M) \\ = & \sum_{m=1}^M \int_{I_m} |x-q_m|^2 f(x)dx - \sum_{m=1}^M \int_{I_m^\star} |x-q_m|^2 f(x)dx \\ = & \sum_{m=1}^M \int_{I_m} |x-q_m|^2 f(x)dx - \sum_{m=1}^M \int_{I_m^\star} \left(\min_{k=1,\ldots,M} |x-q_k|^2\right) f(x)dx \\ = & \sum_{m=1}^M \int_{I_m} |x-q_m|^2 f(x)dx - \int_A^B \left(\min_{k=1,\ldots,M} |x-q_k|^2\right) f(x)dx \\ = & \sum_{m=1}^M \int_{I_m} |x-q_m|^2 f(x)dx - \sum_{m=1}^M \int_{I_m} \left(\min_{k=1,\ldots,M} |x-q_k|^2\right) f(x)dx \\ = & \sum_{m=1}^M \left(\int_{I_m} |x-q_m|^2 f(x)dx - \int_{I_m} \left(\min_{k,\ldots,M} |x-q_k|^2\right) f(x)dx \right) \\ = & \sum_{m=1}^M \int_{I_m} \left(|x-q_m|^2 - \left(\min_{k=1,\ldots,M} |x-q_k|^2\right)\right) f(x)dx \\ \geq & \ 0. \end{aligned} \tag{8}$$

Thus, given the representation levels $q_1,\ldots,q_M$, the cells $I_1^\star,\ldots,I_M^\star$ are given by

$$I_m^\star = (A_m^\star, B_m^\star], \quad m=1,\ldots,M$$

with

$$A_1^\star = A, \quad A_m^\star = \frac{q_{m-1}+q_m}{2}, \quad m=2,\ldots,M.$$

An iterative process

Imagine that you need to minimize the function $H : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ where p and q are positive integers. Although this is a complicated function, assume that it is fairly easy to perform the following two minimizations:

- For each $\boldsymbol{x}$ in $\mathbb{R}^p$,

$$\text{Minimize } H(\boldsymbol{x},\boldsymbol{y}) \text{ with respect to } \boldsymbol{y} \in \mathbb{R}^q$$

with solution $\boldsymbol{y}^\star(\boldsymbol{x})$, i.e.,

$$H(\boldsymbol{x}, \boldsymbol{y}^\star(\boldsymbol{x})), \leq H(\boldsymbol{x}, \boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^q.$$

- For each $\boldsymbol{y}$ in $\mathbb{R}^q$,

  Minimize $H(\boldsymbol{x}, \boldsymbol{y})$ with respect to $\boldsymbol{x} \in \mathbb{R}^p$

  with solution $\boldsymbol{x}^\star(\boldsymbol{y})$, i.e.,

$$H(\boldsymbol{x}^\star(\boldsymbol{y}), \boldsymbol{y}) \leq H(\boldsymbol{x}, \boldsymbol{y}), \quad \boldsymbol{x} \in \mathbb{R}^p.$$

On the basis of this information the following two-step *iterative* algorithm suggests itself very naturally:

Pick $\boldsymbol{x}_1$ in $\mathbb{R}^p$ and set $\boldsymbol{y}_1 = \boldsymbol{y}^\star(\boldsymbol{x}_1)$, so that

$$H(\boldsymbol{x}_1, \boldsymbol{y}_1) \leq H(\boldsymbol{x}_1, \boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^q.$$

Next, with $\boldsymbol{x}_2 = \boldsymbol{x}^\star(\boldsymbol{y}_1)$, it is plain that

$$H(\boldsymbol{x}_2, \boldsymbol{y}_1) \leq H(\boldsymbol{x}, \boldsymbol{y}_1), \quad \boldsymbol{x} \in \mathbb{R}^p,$$

whence

$$H(\boldsymbol{x}_2, \boldsymbol{y}_1) \leq H(\boldsymbol{x}_1, \boldsymbol{y}_1).$$

Similarly, if we set $\boldsymbol{y}_2 = \boldsymbol{y}^\star(\boldsymbol{x}_2)$, then

$$H(\boldsymbol{x}_2, \boldsymbol{y}_2) \leq H(\boldsymbol{x}_2, \boldsymbol{y}_1).$$

Repeating this procedure yields a sequence $\{(\boldsymbol{x}_n, \boldsymbol{y}_n),\ n = 1, 2, \ldots\}$ in $\mathbb{R}^p \times \boldsymbol{R}^q$ through

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}^\star(\boldsymbol{y}_n) \quad \text{and} \quad \boldsymbol{y}_{n+1} = \boldsymbol{y}^\star(\boldsymbol{x}_{n+1}).$$

By construction we get

$$H(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n) \leq H(\boldsymbol{x}_n, \boldsymbol{y}_n)$$

and

$$H(\boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1}) \leq H(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n)$$

for all $n = 1, 2, \ldots$. It follows that the sequence $\{H(\boldsymbol{x}_n, \boldsymbol{y}_n),\ n = 1, 2, \ldots\}$ is non-decreasing, hence its limit

$$L = \lim_{n\to\infty} H(\boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1})$$

always exists. A number of natural questions arise:

1. Is it the case that

$$L = \min\left(H(\boldsymbol{x}, \boldsymbol{y}), \quad \boldsymbol{x} \in \mathbb{R}^p,\ \boldsymbol{y} \in \mathbb{R}^q\right).$$

2. Is there a point $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ in $\mathbb{R}^p \times \mathbb{R}^q$ such that

$$\lim_{n\to\infty} (\boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1}) = (\boldsymbol{x}^\star, \boldsymbol{y}^\star).$$

3. Is it the case then that

$$H(\boldsymbol{x}^\star, \boldsymbol{y}^\star) = L.$$

Thes developments of the two previous sections suggest an *iterative* approach to solving the quantization problem as we identify

$$\boldsymbol{x} \leftarrow (I_1, \ldots, I_M),$$

$$\boldsymbol{y} \leftarrow (q_1, \ldots, q_M)$$

and

$$H(\boldsymbol{x}, \boldsymbol{y}) \leftarrow \Phi_F(I_1, \ldots, I_m; q_1 1, \ldots, q_M).$$

Uniformly distributed samples revisited

The rv X is uniformly distributed on I if its probability density function f is given by

$$f(x) = \begin{cases} 0 & \text{if } x \le A \\ \frac{1}{B-A} & \text{if } A < x \le B \\ 0 & \text{if } B < x. \end{cases}$$

We shall apply the iteration process outlined above:

Start with the cells $I+1,\ldots,I_M$ with

$$I_m = (A_m, B_m], \quad m = 1,2,\ldots,M.$$

In that case, the best representation levels take a particularly simple form: For each $m=1,\ldots,M$, we have

$$\begin{aligned} q^\star_m &= \frac{\int_{A_m}^{B_m} \frac{x}{B-A} dx}{\int_{A_m}^{B_m} \frac{1}{B-A} dx} \\ &= \frac{\int_{A_m}^{B_m} x dx}{\int_{A_m}^{B_m} dx} \\ &= \frac{B_m^2 - A_m^2}{2(B_m - A_m)} \\ &= \frac{A_m + B_m}{2} \end{aligned} \tag{9}$$

so that $q^\star_m$ is the mid-point of the interval I_m.

Next, consider the representation levels $q_1,\ldots,q_m$ with

$$A < q_1, \ldots < q_M < B.$$

As indicated earlier, the best corresponding cells are the intervals $I^\star_1,\ldots,I^\star_m$ of the form

$$I^\star_m = (A^\star_m, B^\star_m], \quad m = 1,\ldots,M$$

with

$$A^\star_1 = A, \quad A^\star_m = \frac{q_{m-1}+q_m}{2}, \quad m = 2,\ldots,M.$$

A classical calculation of the signal-to-quantization-noise ratio

Consider the *uniform* quantizer $Q^u = (I^u_1,\ldots,I^u_M; q^u_1,\ldots,q^u_M)$. For each $m = 1,\ldots,M$, whenever x lies in the interval I^u_m, we have

$$\varepsilon(Q^u;x) = Q^u(x) - x = q^u_m - x$$

an q^u_m being the midpoint of the interval I^u_m, it follows that

$$|x - q^u_m| \le \frac{B-A}{2M}.$$

As a result, the rv $X - Q^u(X)$ takes values in the symmetric interval

$$J := \left[-\frac{B-A}{2M}, \frac{B-A}{2M} \right].$$

It is easy to see that if the density f is sufficiently smooth and M is sufficiently large,[2] then the probability distribution of the rv $X - Q^u(X)$ is well approximated by the uniform distribution on the interval J. Thus,

$$\begin{aligned} \mathbb{E}\left[|\varepsilon(Q^u; X)|^2 \right] &= \int_{-\frac{B-A}{2M}}^{\frac{B-A}{2M}} t^2 f_{\varepsilon_{Q^u}(X)}(t) dt \\ &\simeq \int_{-\frac{B-A}{2M}}^{\frac{B-A}{2M}} \frac{t^2}{\frac{B-A}{M}} dt \\ &= \frac{M}{B-A} \cdot \left[\frac{t^3}{3} \right]_{-\frac{B-A}{2M}}^{\frac{B-A}{2M}} \\ &= \frac{2M}{3(B-A)} \cdot \left(\frac{B-A}{2M} \right)^3 \end{aligned} \tag{10}$$

so that

$$\mathbb{E}\left[|\varepsilon(Q^u; X)|^2 \right] \simeq \frac{1}{12} \cdot \left(\frac{B-A}{M} \right)^2.$$

Finally,

$$\begin{aligned} \text{SQNR}(Q^u; X) &= \frac{\mathbb{E}\left[X^2 \right]}{\mathbb{E}\left[|\varepsilon(Q^u; X)|^2 \right]} \\ &\simeq 12 \frac{\mathbb{E}\left[X^2 \right]}{(B-A)^2} \cdot M^2. \end{aligned} \tag{11}$$

It is customary to write

$$M = 2^R$$

[2] As the calculations given at the end of the writeup show, these conditions have to be read to saying that the staircase approximation of f anchored at the points $A + m\frac{B-A}{M}$, $m = 0, 1, \ldots, M - 1$ is indeed a (reasonably) good approximation of f.

where R is the size of the binary representation of M. With this notation we get

$$\mathrm{SQNR}(Q^u; X) \simeq C(X) \cdot 2^{2R}$$

where the first factor

$$C(X) = 12 \frac{\mathbb{E}\left[X^2\right]}{(B-A)^2}$$

is determined only by the source, while the second factor 2^{2R} expresses the coarseness of the approximation of the quantizer. Thus,

$$\begin{aligned} \mathrm{SQNR}(Q^u; X)_{dB} &\simeq 10 \log_{10} C(X) + 20R \cdot \log_{10} 2 \ (dB) \\ &= 10 \log_{10} C(X) + 6.02 \cdot R \ (dB) \end{aligned} \tag{12}$$

as we recall that $\log_{10} 2 = 0.30102999 \ldots$. Adding one extra bit means two levels with the net result that the SQNR increases by $.02$ dB.

Companding – Non-uniform quantizers through composition

With $\tilde{A} < \tilde{B}$, define the interval $\tilde{I} = (\tilde{A}, \tilde{B}]$. Assume given a continuous mapping $\Phi : I \to \tilde{I}$ which is *strictly monotone increasing* with

$$\tilde{A} = \Phi(A) \quad \text{and} \quad \tilde{B} = \Phi(B).$$

Thus, Φ puts the intervals I and $\tilde{I}$ into *one-to-one* correspondence. The case of interest is when Φ is *non*-linear.

Let X denote a rv with a non-uniform distribution on the interval I. With $\widetilde{X} := \Phi(X)$, the rv $\widetilde{X}$ is distributed on the interval $\tilde{I}$. We shall quantize its samples by means of the uniform quantizer for the interval $\widetilde{I}$, namely

$$\widetilde{Q}^u \equiv (\widetilde{I}_1^u, \ldots, \widetilde{I}_M^u; \widetilde{q}_1^u, \ldots, \widetilde{q}_M^u)$$

with cells

$$\widetilde{I}_m^u = (\tilde{A}_m^u, \widetilde{B}_m^u], \quad m = 1, \ldots, M$$

and representation levels

$$\widetilde{q}_m^u = \frac{\tilde{A}_m^u + \widetilde{B}_m^u}{2}, \quad m = 1, \ldots, M$$

where

$$\tilde{A}_m^u = \tilde{A} + (m-1) \cdot \frac{\tilde{B} - \tilde{A}}{M}.$$

This uniform quantizer $\widetilde{Q}^u$, through the intermediary of Φ, produces a *non-uniform* quantizer Q for X by setting

$$Q(x) := \Phi^{-1}\left(\widetilde{Q}^u(\Phi(x))\right), \quad x \in I.$$

This procedure is known as *companding*, an abbreviation for *comp*ressing followed by ex*panding*.

It is easy to check that this procedure indeed defines a quantizer Q for the interval I with cells $I_1, \ldots, I_M$ and representation levels $q_1, \ldots, q_M$ given by

$$I_m := \Phi^{-1}(\widetilde{I}^u_m) \quad \text{and} \quad q_m := \Phi^{-1}(\widetilde{q}^u_m), \quad m = 1, \ldots, M.$$

The interval I_m is of the form $(A_m, B_m]$ with endpoints

$$A_m = \Phi^{-1}(\widetilde{A}^u_m) \quad \text{and} \quad B_m = \Phi^{-1}(\widetilde{B}^u_m).$$

In short,

$$Q(x) = \Phi^{-1}(\widetilde{q}^u_m), \quad x \in I_m,\ m = 1, \ldots, M.$$

The function Φ is selected so as to capture key features of the distribution of X, e.g., its skewness. This is done by trial and error, by using functions that belong to well structured classes of functions. This approach obviates the need to solve the quantization problem, usually a difficult task, either directly or through the iterative procedure outlined earlier. While companding may yield a sub-optimal quantizer (with respect to the mean-square distortion metric used earlier), its robustness and ease of implementation are traded for acceptable performance.

Fix $m = 1, \ldots, M$. By construction, we have

$$\begin{aligned} \widetilde{B}^u_m - \widetilde{A}^u_m &= \Phi(B_m) - \Phi(A_m) \\ &= \int_{I_m} \Phi'(x)dx \end{aligned} \tag{13}$$

under weak differentiability assumptions. Now note that

$$\widetilde{B}^u_m - \widetilde{A}^u_m = \frac{\widetilde{B} - \widetilde{A}}{M}$$

while

$$\int_{I_m} \Phi'(x)dx \simeq (B_m - A_m)\Phi'(q_m).$$

Comparing we see that

$$\frac{\widetilde{B} - \widetilde{A}}{M} \simeq (B_m - A_m)\Phi'(q_m)$$

so that

$$B_m - A_m \simeq \frac{\widetilde{B} - \widetilde{A}}{M\Phi'(q_m)}.$$

The μ and A-laws

In practice the interval $I = (A, B]$ is symmetric with respect to the origin with $A = -B$ for $B > 0$, the interval $\widetilde{I}$ coincides with I, and the compressor is an *odd* strictly increasing and continuous function $\Phi : I \to I$ with

$$\Phi(-x) = -\Phi(x), \quad |x| \leq B$$

and

$$\Phi(\pm B) = \pm B.$$

Companding has been deployed in telephone networks as part of the PCM format. Two standards have emerged: The μ-law is used in the U.S, Canada and Japan, while the A-law has been adopted in Europe. They are briefly discussed below.

With $\mu > 0$, the μ-law corresponds to the mapping $\Phi_\mu : [-B, B] \to [-B, B]$ given by

(14) $$\Phi_\mu(x) = B \cdot \frac{\ln\left(1 + \mu\frac{|x|}{B}\right)}{\ln(1+\mu)} \cdot \mathrm{sgn}(x), \quad |x| \leq B$$

For $\mu = 0$, we find $\Phi_\mu(x) = x$ on the interval $[-B, B]$ and companding reduces to uniform quantization on I.

With $A > 1$, the A-law is defined through the mapping $\Phi_A : [-B, B] \to [-B, B]$ given by

(15) $$\Phi_A(x) := \begin{cases} \frac{A}{1+\ln A} \cdot |x| \cdot \mathrm{sgn}(x) & \text{if } \frac{|x|}{B} \leq \frac{1}{A} \\ B\frac{1+\ln\left(A\frac{|x|}{B}\right)}{1+\ln A} \cdot \mathrm{sgn}(x) & \text{if } \frac{1}{A} \leq \frac{|x|}{B} \leq 1 \end{cases}$$

with $A > 1$. The value $A = 1$ yields $\Phi_A(x) = x$ on the interval $[-B, B]$, in which case companding reduces to uniform quantization on I.

Approximating the probability density function of the quantization noise under a uniform quantizer

Pick t in the interval J where

$$J = \left[-\frac{B-A}{2M}, \frac{B-A}{2M} \right].$$

By standard probabilistic arguments,

$$\begin{aligned}
\mathbb{P}\left[\varepsilon(Q^u; X) \le t\right] &= \sum_{m=1}^{M} \mathbb{P}\left[X \in I^u_m, \varepsilon(Q^u; X) \le t\right] \\
&= \sum_{m=1}^{M} \mathbb{P}\left[X \in I^u_m, Q^u(X) - X \le t\right] \\
&= \sum_{m=1}^{M} \mathbb{P}\left[X \in I^u_m, q^u_m - X \le t\right] \\
&= \sum_{m=1}^{M} \mathbb{P}\left[X \in I^u_m, q^u_m - t \le X\right] \\
&= \sum_{m=1}^{M} \mathbb{P}\left[A^u_m < X \le B^u_m, q^u_m - t \le X\right] \\
&= \sum_{m=1}^{M} \mathbb{P}\left[q^u_m - t \le X \le B^u_m\right] \\
&= \sum_{m=1}^{M} \int_{q^u_m - t}^{B^u_m} f(x)dx \qquad (16)
\end{aligned}$$

as we have used the fact that q^u_m is the midpoint between A^u_m and B^u_m (which are themselves $\frac{B-A}{M}$ apart of each other), so that

$$A^u_m < q^u_m - t \le B^u_m, \quad t \in J.$$

If the probability density function f of X is sufficiently smooth and M is sufficiently large, then the approximation

$$f(x) \simeq f(q^u_m), \quad x \in I^u_m, \ m = 1, \ldots, M$$

is likely to hold since each of the intervals $I_1^u, \ldots, I_M^u$ is small. Reporting this fact into the result of the earlier calculations we get

$$
\begin{aligned}
\mathbb{P}\left[\varepsilon(Q^u;X) \leq t\right] &= \sum_{m=1}^{M} \int_{q_m^u - t}^{B_m^u} f(x)dx \\
&\simeq \sum_{m=1}^{M} \int_{q_m^u - t}^{B_m^u} f(q_m^u)dx \\
&= \sum_{m=1}^{M} f(q_m^u)\left(B_m^u - (q_m^u - t)\right) \\
&= \sum_{m=1}^{M} f(q_m^u)\left(\frac{B-A}{2M} + t\right) \qquad (17)
\end{aligned}
$$

since

$$
\begin{aligned}
B_m^u - q_m^u &= \left(A + m \cdot \frac{B-A}{M}\right) - \left(A + \frac{2m-1}{2} \cdot \frac{B-A}{M}\right) \\
&= \frac{B-A}{2M}. \qquad (18)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{P}\left[\varepsilon(Q^u;X) \leq t\right] &\simeq \left(\sum_{m=1}^{M} f(q_m^u)\right) \cdot \left(\frac{B-A}{2M} + t\right) \\
&= \left(\sum_{m=1}^{M} f(q_m^u)\frac{B-A}{M}\right) \cdot \left(\frac{1}{2} + \frac{M}{B-A} \cdot t\right) \\
&\simeq \frac{1}{2} + \frac{M}{B-A} \cdot t, \quad t \in J. \qquad (19)
\end{aligned}
$$

The last step leading to (19) relies on the approximation argument used earlier but in the following reversed way: We see that

$$
\begin{aligned}
\sum_{m=1}^{M} f(q_m^u)\frac{B-A}{M} &= \sum_{m=1}^{M} \int_{I_m^u} f(q_m^u)dx \\
&\simeq \sum_{m=1}^{M} \int_{I_m^u} f(x)dx
\end{aligned}
$$

$$= \int_I f(x)dx$$

(20) $$= 1$$

since a probability density function integrates to unity. It is now straightforward to see that (19) is the probability distribution function of a rv which is uniformly distributed on J.

SQNR under companding

Let Q denote the non-uniform quantizer obtained by companding through the compressor $\Phi : (A, B] \to (\tilde{A}, \tilde{B}]$.

$$\begin{aligned}
\mathbb{E}\left[|\varepsilon(Q;X)^2|\right] &= \sum_{m=1}^{M} \int_{I_m} (Q(x)-x)^2 f(x)dx \\
&= \sum_{m=1}^{M} \int_{I_m} (q_m - x)^2 f(x)dx \\
&\simeq \sum_{m=1}^{M} f(q_m) \int_{I_m} (q_m - x)^2 dx
\end{aligned}$$

(21) $$= \sum_{m=1}^{M} f(q_m)$$

since

$$\begin{aligned}
\int_{I_m} (q_m - x)^2 dx &= \int_{A_m}^{B_m} (q_m - x)^2 dx \\
&= \left[\frac{-(q_m - x)^3}{3}\right]_{A_m}^{B_m}
\end{aligned}$$

(22) $$= \frac{1}{3}\left((q_m - A_m)^3 - (q_m - B_m)^3\right).$$