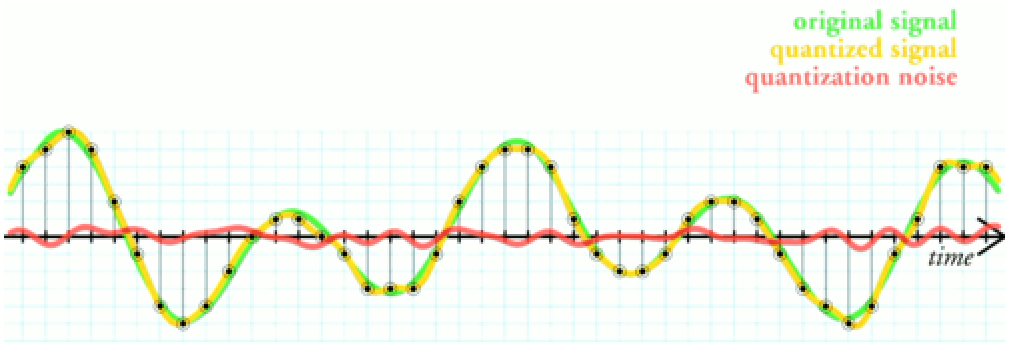
Quantization (signal processing)

Quantization, in mathematics and digital signal processing, is the process of mapping input values from a large set (often a continuous set) to output values in a (countable) smaller set. Rounding and truncation are typical examples of quantization processes. Quantization is involved to some degree in nearly all digital signal processing, as the process of representing a signal in digital form ordinarily involves rounding. Quantization also forms the core of essentially all lossy compression algorithms.

The difference between an input value and its quantized value (such as round-off error) is referred to as **quantization error**. A device or algorithmic function that performs quantization is called a **quantizer**. An analog-to-digital converter is an example of a quantizer.



The simplest way to quantize a signal is to choose the digital amplitude value closest to the original analog amplitude. This example shows the original analog signal (green), the quantized signal (black dots), the signal reconstructed from the quantized signal (yellow) and the difference between the original signal and the reconstructed signal (red). The difference between the original signal and the reconstructed signal is the quantization error and, in this simple quantization scheme, is a deterministic function of the input signal.

Contents

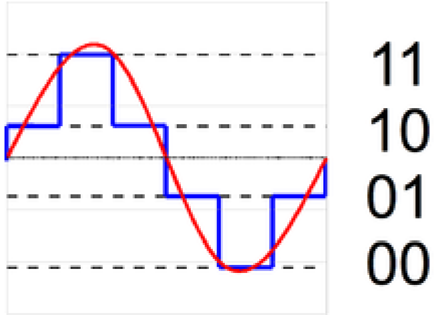
Basic properties of quantization

Because quantization is a many-to-few mapping, it is an inherently non-linear and irreversible process (i.e., because the same output value is shared by multiple input values, it is impossible in general to recover the exact input value when given only the output value).

The set of possible input values may be infinitely large, and may possibly be continuous and therefore uncountable (such as the set of all real numbers, or all real numbers within some limited range). The set of possible output values may be finite or countably infinite. The input and output sets involved in quantization can be defined in a rather general way. For example, *vector quantization* is the application of quantization to multi-dimensional (vector-valued) input data.[1]

Basic types of quantization

Analog-to-digital converter (ADC)



2-bit resolution with four levels of quantization compared to analog.[2]

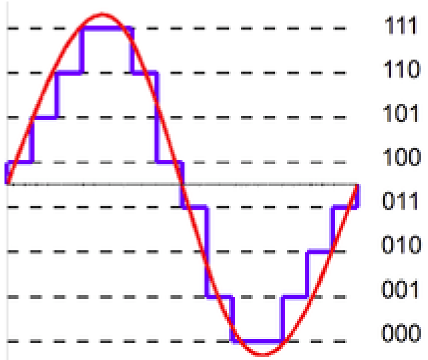
Outside the realm of signal processing, this category may simply be called *rounding* or *scalar quantization*. An ADC can be modeled as two processes: sampling and **quantization**. Sampling converts a voltage signal (function of time) into a discrete-time signal (sequence of real numbers). Quantization replaces each real number with an approximation from a finite set of discrete values (**levels**), which is necessary for storage and processing by numerical methods. Most commonly, these discrete values are represented as fixed-point words (either proportional to the waveform values or companded) or floating-point words. Common word-lengths are 8-bit (256 levels), 16-bit (65,536 levels), 32-bit (4.3 billion levels), and so on, though any number of quantization levels is possible (not just powers of two). Quantizing a sequence of numbers produces a sequence of quantization errors which is sometimes modeled as an additive random signal called **quantization noise** because of its stochastic behavior. The more levels a quantizer uses, the lower is its quantization noise power.

In general, both ADC processes lose some information. So discrete-valued signals are only an approximation of the continuous-valued discrete-time signal, which is itself only an approximation of the original continuous-valued continuous-time signal. But both types of approximation errors can, in theory, be made arbitrarily small by good design.

Rate–distortion optimization

Rate–distortion optimized quantization is encountered in source coding for "lossy" data compression algorithms, where the purpose is to manage distortion within the limits of the bit rate supported by a communication channel or storage medium. In this second setting, the amount of introduced distortion may be managed carefully by sophisticated techniques, and introducing some significant amount of distortion may be unavoidable. A quantizer designed for this purpose may be quite different and more elaborate in design than an ordinary rounding operation. It is in this domain that substantial rate–distortion theory analysis is likely to be applied. However, the same concepts actually apply in both use cases.

The analysis of quantization involves studying the amount of data (typically measured in digits or bits or bit *rate*) that is used to represent the output of the quantizer, and studying the loss of precision that is introduced by the quantization process (which is referred to as the *distortion*). The general field of such study of rate and distortion is known as *rate–distortion theory*.



3-bit resolution with eight levels.

Rounding example

As an example, rounding a real number x to the nearest integer value forms a very basic type of quantizer – a *uniform* one. A typical (*mid-tread*) uniform quantizer with a quantization *step size* equal to some value Δ can be expressed as

$$Q(x) = \Delta \cdot \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor = \Delta \cdot \operatorname{floor}\left(\frac{x}{\Delta} + \frac{1}{2} \right),$$

where the notation $\lfloor\ \rfloor$ or $\operatorname{floor}(\ )$ depicts the floor function. For simple rounding to the nearest integer, the step size Δ is equal to 1. With $\Delta = 1$ or with Δ equal to any other integer value, this quantizer has real-valued inputs and integer-valued outputs, although this property is not a necessity – a quantizer may also have an integer input domain and may also have non-integer output values. The essential property of a quantizer is that it has a countable set of possible output values that has fewer members than the set of possible input values. The members of the set of output values may have integer, rational, or real values (or even other possible values as well, in general – such as vector values or complex numbers).

When the quantization step size is small (relative to the variation in the signal being measured), it is relatively simple to show[3][4][5][6][7][8] that the mean squared error produced by such a rounding operation will be approximately $\Delta^2/12$. Mean squared error is also called the quantization **noise power**. Adding one bit to the quantizer halves the value of Δ, which reduces the noise power by the factor ¼. In terms of decibels, the noise power change is $10 \cdot \log_{10}\left(\frac{1}{4}\right) \approx -6\ \mathrm{dB}$.

Because the set of possible output values of a quantizer is countable, any quantizer can be decomposed into two distinct stages, which can be referred to as the *classification* stage (or *forward quantization* stage) and the *reconstruction* stage (or *inverse quantization* stage), where the classification stage maps the input value to an integer *quantization index* k and the reconstruction stage maps the index k to the *reconstruction value* y_k that is the output approximation of the input value. For the example uniform quantizer described above, the forward quantization stage can be expressed as

$$k = \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor,$$

and the reconstruction stage for this example quantizer is simply

$$y_k = k \cdot \Delta.$$

This decomposition is useful for the design and analysis of quantization behavior, and it illustrates how the quantized data can be communicated over a communication channel – a *source encoder* can perform the forward quantization stage and send the index information through a communication channel (possibly applying entropy coding techniques to the quantization indices), and a *decoder* can perform the reconstruction stage to produce the output approximation of the original input data. In more elaborate quantization designs, both the forward and inverse quantization stages may be substantially more complex. In general, the forward quantization stage may use any function that maps the input data to the integer space of the quantization index data, and the inverse quantization stage can conceptually (or literally) be a table look-up operation to map each quantization index to a corresponding reconstruction value. This two-stage decomposition applies equally well to vector as well as scalar quantizers.

Mid-riser and mid-tread uniform quantizers

Most uniform quantizers for signed input data can be classified as being of one of two types: **mid-riser** and **mid-tread**. The terminology is based on what happens in the region around the value 0, and uses the analogy of viewing the input-output function of the quantizer as a stairway. Mid-tread quantizers have a zero-valued reconstruction level (corresponding to a *tread* of a stairway), while mid-riser quantizers have a zero-valued classification threshold (corresponding to a *riser* of a stairway).[9]

The formulas for mid-tread uniform quantization are provided in the previous section.

The input-output formula for a mid-riser uniform quantizer is given by:

$$Q(x) = \Delta \cdot \left(\left\lfloor \frac{x}{\Delta} \right\rfloor + \frac{1}{2} \right),$$

where the classification rule is given by

$$k = \left\lfloor \frac{x}{\Delta} \right\rfloor$$

and the reconstruction rule is

$$y_k = \Delta \cdot \left(k + \tfrac{1}{2} \right).$$

Note that mid-riser uniform quantizers do not have a zero output value – their minimum output magnitude is half the step size. When the input data can be modeled as a random variable with a probability density function (pdf) that is smooth and symmetric around zero, mid-riser quantizers also always produce an output *entropy* of at least 1 bit per sample.

In contrast, mid-tread quantizers do have a zero output level, and can reach arbitrarily low bit rates per sample for input distributions that are symmetric and taper off at higher magnitudes. For some applications, having a zero output signal representation or supporting low output entropy may be a necessity. In such cases, using a mid-tread uniform quantizer may be appropriate while using a mid-riser one would not be.

In general, a mid-riser or mid-tread quantizer may not actually be a *uniform* quantizer – i.e., the size of the quantizer's classification intervals may not all be the same, or the spacing between its possible output values may not all be the same. The distinguishing characteristic of a mid-riser quantizer is that it has a classification threshold value that is exactly zero, and the distinguishing characteristic of a mid-tread quantizer is that is it has a reconstruction value that is exactly zero.[9]

Dead-zone quantizers

Another name for a mid-tread quantizer with symmetric behavior around 0 is **dead-zone quantizer**, and the classification region around the zero output value of such a quantizer is referred to as the *dead zone* or *deadband*. The dead zone can sometimes serve the same purpose as a noise gate or squelch function. Especially for compression applications, the dead-zone may be given a different width than that for the other steps. For an otherwise-uniform quantizer, the dead-zone width can be set to any value w by using the forward quantization rule[10][11][12]

$$k = \operatorname{sgn}(x) \cdot \max\left(0, \left\lfloor \frac{|x| - w/2}{\Delta} + 1 \right\rfloor\right),$$

where the function $\operatorname{sgn}(\ )$ is the sign function (also known as the *signum* function). The general reconstruction rule for such a dead-zone quantizer is given by

$$y_k = \operatorname{sgn}(k) \cdot \left(\frac{w}{2} + \Delta \cdot (|k| - 1 + r_k)\right),$$

where r_k is a reconstruction offset value in the range of 0 to 1 as a fraction of the step size. Ordinarily, $0 \leq r_k \leq \frac{1}{2}$ when quantizing input data with a typical pdf that is symmetric around zero and reaches its peak value at zero (such as a Gaussian, Laplacian, or Generalized Gaussian pdf). Although r_k may depend on k in general, and can be chosen to fulfill the optimality condition described below, it is often simply set to a constant, such as $\frac{1}{2}$. (Note that in this definition, $y_0 = 0$ due to the definition of the $\operatorname{sgn}(\ )$ function, so r_0 has no effect.)

A very commonly used special case (e.g., the scheme typically used in financial accounting and elementary mathematics) is to set $w = \Delta$ and $r_k = \frac{1}{2}$ for all k.

Granular distortion and overload distortion

Often the design of a quantizer involves supporting only a limited range of possible output values and performing clipping to limit the output to this range whenever the input exceeds the supported range. The error introduced by this clipping is referred to as *overload* distortion. Within the extreme limits of the supported range, the amount of spacing between the selectable output values of a quantizer is referred to as its *granularity*, and the error introduced by this spacing is referred to as *granular* distortion. It is common for the design of a quantizer to involve determining the proper balance between granular distortion and overload distortion. For a given supported number of possible output values, reducing the average granular distortion may involve increasing the average overload distortion, and vice versa. A technique for controlling the amplitude of the signal (or, equivalently, the quantization step size Δ) to achieve the appropriate balance is the use of *automatic gain control* (AGC). However, in some quantizer designs, the concepts of granular error and overload error may not apply (e.g., for a quantizer with a limited range of input data or with a countably infinite set of selectable output values).

The additive noise model for quantization error

A common assumption for the analysis of quantization error is that it affects a signal processing system in a similar manner to that of additive white noise – having negligible correlation with the signal and an approximately flat power spectral density.[4][8][13][14] The additive noise model is commonly used for the analysis of quantization error effects in digital filtering systems, and it can be very useful in such analysis. It has been shown to be a valid model in cases of high resolution quantization (small Δ relative to the signal strength) with smooth probability density functions.[4][15] However, additive noise behaviour is not always a valid assumption, and care should be taken to avoid assuming that this model always applies. In actuality, the quantization error (for quantizers defined as described here) is deterministically related to the signal rather than being independent of it.[8] Thus, periodic signals can create periodic quantization noise. And in some cases it can even cause limit cycles to appear in digital signal processing systems.[14]

One way to ensure effective independence of the quantization error from the source signal is to perform *dithered quantization* (sometimes with *noise shaping*), which involves adding random (or pseudo-random) noise to the signal prior to quantization.[8][14] This can sometimes be beneficial for such purposes as improving the subjective quality of the result, however it can increase the total quantity of error introduced by the quantization process.

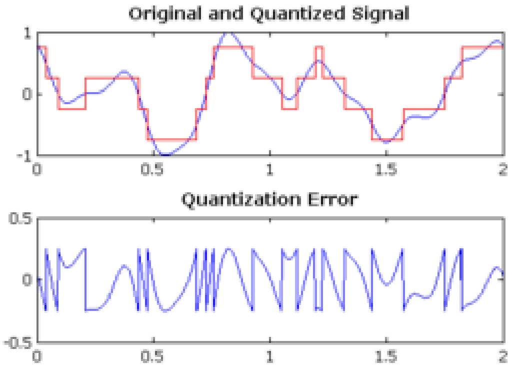
Quantization error models

In the typical case, the original signal is much larger than one least significant bit (LSB). When this is the case, the quantization error is not significantly correlated with the signal, and has an approximately uniform distribution. In the rounding case, the quantization error has a mean of zero and the RMS value is the standard deviation of this distribution, given by $\frac{1}{\sqrt{12}}\mathrm{LSB} \approx 0.289\,\mathrm{LSB}$. In the truncation case the error has a non-zero mean of $\frac{1}{2}\mathrm{LSB}$ and the RMS value is $\frac{1}{\sqrt{3}}\mathrm{LSB}$. In either case, the standard deviation, as a percentage of the full signal range, changes by a factor of 2 for each 1-bit change in the number of quantizer bits. The potential signal-to-quantization-noise power ratio therefore changes by 4, or $10\cdot\log_{10}(4) = 6.02$ *decibels per bit*.

At lower amplitudes the quantization error becomes dependent on the input signal, resulting in distortion. This distortion is created after the anti-aliasing filter, and if these distortions are above 1/2 the sample rate they will alias back into the band of interest. In order to make the quantization error independent of the input signal, noise with an amplitude of 2 least significant bits is added to the signal. This slightly reduces signal to noise ratio, but, ideally, completely eliminates the distortion. It is known as dither.

Quantization noise model

Quantization noise is a model of quantization error introduced by quantization in the analog-to-digital conversion (ADC) in telecommunication systems and signal processing. It is a rounding error between the analog input voltage to the ADC and the output digitized value. The noise is non-linear and signal-dependent. It can be modelled in several different ways.



Quantization noise for a 2-bit ADC operating at infinite sample rate. The difference between the blue and red signals in the upper graph is the quantization error, which is "added" to the quantized signal and is the source of noise.

In an ideal analog-to-digital converter, where the quantization error is uniformly distributed between −1/2 LSB and +1/2 LSB, and the signal has a uniform distribution covering all quantization levels, the Signal-to-quantization-noise ratio (SQNR) can be calculated from

$$\mathrm{SQNR} = 20\log_{10}(2^Q) \approx 6.02 \cdot Q\ \mathrm{dB}$$

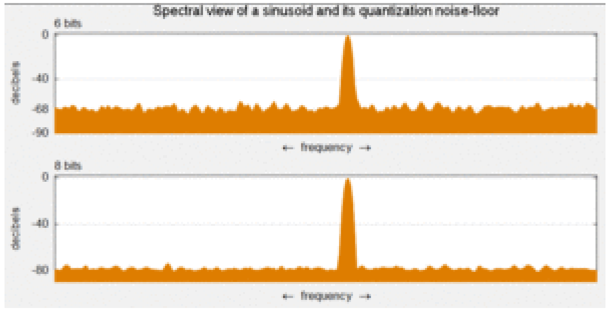
Where Q is the number of quantization bits.

The most common test signals that fulfill this are full amplitude triangle waves and sawtooth waves.

For example, a 16-bit ADC has a maximum signal-to-noise ratio of 6.02 × 16 = 96.3 dB.

When the input signal is a full-amplitude sine wave the distribution of the signal is no longer uniform, and the corresponding equation is instead

$$\mathrm{SQNR} \approx 1.761 + 6.02 \cdot Q\ \mathrm{dB}$$



Comparison of quantizing a sinusoid to 64 levels (6 bits) and 256 levels (8 bits). The additive noise created by 6-bit quantization is 12 dB greater than the noise created by 8-bit quantization. When the spectral distribution is flat, as in this example, the 12 dB difference manifests as a measurable difference in the noise floors.

Here, the quantization noise is once again *assumed* to be uniformly distributed. When the input signal has a high amplitude and a wide frequency spectrum this is the case.[16] In this case a 16-bit ADC has a maximum signal-to-noise ratio of 98.09 dB. The 1.761 difference in signal-to-noise only occurs due to the signal being a full-scale sine wave instead of a triangle/sawtooth.

Quantization noise power can be derived from

$$\mathrm{N} = \frac{(\delta \mathrm{v})^2}{12}\mathrm{W}$$

where $\delta\mathrm{v}$ is the voltage of the level.

(Typical real-life values are worse than this theoretical minimum, due to the addition of dither to reduce the objectionable effects of quantization, and to imperfections of the ADC circuitry. Also see noise shaping.)

For complex signals in high-resolution ADCs this is an accurate model. For low-resolution ADCs, low-level signals in high-resolution ADCs, and for simple waveforms the quantization noise is not uniformly distributed, making this model inaccurate.[17] In these cases the quantization noise distribution is strongly affected by the exact amplitude of the signal.

The calculations above, however, assume a completely filled input channel. If this is not the case - if the input signal is small - the relative quantization distortion can be very large. To circumvent this issue, analog compressors and expanders can be used, but these introduce large amounts of distortion as well, especially if the compressor does not match the expander. The application of such compressors and expanders is also known as companding.

Rate–distortion quantizer design

A scalar quantizer, which performs a quantization operation, can ordinarily be decomposed into two stages:

- **Classification:** A process that classifies the input signal range into M non-overlapping **intervals** $\{I_k\}_{k=1}^{M}$, by defining $M-1$ **boundary (decision)** values $\{b_k\}_{k=1}^{M-1}$, such that $I_k = [b_{k-1}~,~b_k)$ for $k = 1, 2, \ldots, M$, with the extreme limits defined by $b_0 = -\infty$ and $b_M = \infty$. All the inputs x that fall in a given interval range I_k are associated with the same quantization index k.
- **Reconstruction:** Each interval I_k is represented by a **reconstruction value** y_k which implements the mapping $x \in I_k \Rightarrow y = y_k$.

These two stages together comprise the mathematical operation of $y = Q(x)$.

Entropy coding techniques can be applied to communicate the quantization indices from a source encoder that performs the classification stage to a decoder that performs the reconstruction stage. One way to do this is to associate each quantization index k with a binary codeword c_k. An important consideration is the number of bits used for each codeword, denoted here by $\mathrm{length}(c_k)$.

As a result, the design of an M-level quantizer and an associated set of codewords for communicating its index values requires finding the values of $\{b_k\}_{k=1}^{M-1}$, $\{c_k\}_{k=1}^{M}$ and $\{y_k\}_{k=1}^{M}$ which optimally satisfy a selected set of design constraints such as the **bit rate** R and **distortion** D.

Assuming that an information source S produces random variables X with an associated probability density function $f(x)$, the probability p_k that the random variable falls within a particular quantization interval I_k is given by

$$p_k = P[x \in I_k] = \int_{b_{k-1}}^{b_k} f(x)dx\,.$$

The resulting bit rate R, in units of average bits per quantized value, for this quantizer can be derived as follows:

$$R = \sum_{k=1}^{M} p_k \cdot \mathrm{length}(c_k) = \sum_{k=1}^{M} \mathrm{length}(c_k) \int_{b_{k-1}}^{b_k} f(x)dx.$$

If it is assumed that distortion is measured by mean squared error, the distortion **D**, is given by:

$$D = E[(x - Q(x))^2] = \int_{-\infty}^{\infty} (x - Q(x))^2 f(x)dx = \sum_{k=1}^{M} \int_{b_{k-1}}^{b_k} (x - y_k)^2 f(x)dx.$$

Note that other distortion measures can also be considered, although mean squared error is a popular one.

A key observation is that rate R depends on the decision boundaries $\{b_k\}_{k=1}^{M-1}$ and the codeword lengths $\{\mathrm{length}(c_k)\}_{k=1}^{M}$, whereas the distortion D depends on the decision boundaries $\{b_k\}_{k=1}^{M-1}$ and the reconstruction levels $\{y_k\}_{k=1}^{M}$.

After defining these two performance metrics for the quantizer, a typical Rate–Distortion formulation for a quantizer design problem can be expressed in one of two ways:

1. Given a maximum distortion constraint $D \le D_{\max}$, minimize the bit rate R
2. Given a maximum bit rate constraint $R \le R_{\max}$, minimize the distortion D

Often the solution to these problems can be equivalently (or approximately) expressed and solved by converting the formulation to the unconstrained problem $\min\{D + \lambda \cdot R\}$ where the Lagrange multiplier λ is a non-negative constant that establishes the appropriate balance between rate and distortion. Solving the unconstrained problem is equivalent to finding a point on the convex hull of the family of solutions to an equivalent constrained formulation of the problem. However, finding a solution – especially a closed-form solution – to any of these three problem formulations can be difficult. Solutions that do not require multi-dimensional iterative optimization techniques have been published for only three probability distribution functions: the uniform,[18] exponential,[12] and Laplacian[12] distributions. Iterative optimization approaches can be used to find solutions in other cases.[8][19][20]

Note that the reconstruction values $\{y_k\}_{k=1}^{M}$ affect only the distortion – they do not affect the bit rate – and that each individual y_k makes a separate contribution d_k to the total distortion as shown below:

$$D = \sum_{k=1}^{M} d_k$$

where

$$d_k = \int_{b_{k-1}}^{b_k} (x - y_k)^2 f(x)dx$$

This observation can be used to ease the analysis – given the set of $\{b_k\}_{k=1}^{M-1}$ values, the value of each y_k can be optimized separately to minimize its contribution to the distortion D.

For the mean-square error distortion criterion, it can be easily shown that the optimal set of reconstruction values $\{y_k^*\}_{k=1}^{M}$ is given by setting the reconstruction value y_k within each interval I_k to the conditional expected value (also referred to as the *centroid*) within the interval, as given by:

$$y_k^* = \frac{1}{p_k} \int_{b_{k-1}}^{b_k} x f(x)dx.$$

The use of sufficiently well-designed entropy coding techniques can result in the use of a bit rate that is close to the true information content of the indices $\{k\}_{k=1}^{M}$, such that effectively

$$\text{length}(c_k) \approx -\log_2(p_k)$$

and therefore

$$R = \sum_{k=1}^{M} -p_k \cdot \log_2(p_k).$$

The use of this approximation can allow the entropy coding design problem to be separated from the design of the quantizer itself. Modern entropy coding techniques such as arithmetic coding can achieve bit rates that are very close to the true entropy of a source, given a set of known (or adaptively estimated) probabilities $\{p_k\}_{k=1}^{M}$.

In some designs, rather than optimizing for a particular number of classification regions M, the quantizer design problem may include optimization of the value of M as well. For some probabilistic source models, the best performance may be achieved when M approaches infinity.

Neglecting the entropy constraint: Lloyd–Max quantization

In the above formulation, if the bit rate constraint is neglected by setting λ equal to 0, or equivalently if it is assumed that a fixed-length code (FLC) will be used to represent the quantized data instead of a variable-length code (or some other entropy coding technology such as arithmetic coding that is better than an FLC in the rate–distortion sense), the optimization problem reduces to minimization of distortion D alone.

The indices produced by an M-level quantizer can be coded using a fixed-length code using $R = \lceil \log_2 M \rceil$ bits/symbol. For example when $M =$256 levels, the FLC bit rate R is 8 bits/symbol. For this reason, such a quantizer has sometimes been called an 8-bit quantizer. However using an FLC eliminates the compression improvement that can be obtained by use of better entropy coding.

Assuming an FLC with M levels, the Rate–Distortion minimization problem can be reduced to distortion minimization alone. The reduced problem can be stated as follows: given a source X with pdf $f(x)$ and the constraint that the quantizer must use only M classification regions, find the decision boundaries $\{b_k\}_{k=1}^{M-1}$ and reconstruction levels $\{y_k\}_{k=1}^{M}$ to minimize the resulting distortion

$$D = E[(x - Q(x))^2] = \int_{-\infty}^{\infty} (x - Q(x))^2 f(x)dx = \sum_{k=1}^{M} \int_{b_{k-1}}^{b_k} (x - y_k)^2 f(x)dx = \sum_{k=1}^{M} d_k .$$

Finding an optimal solution to the above problem results in a quantizer sometimes called a MMSQE (minimum mean-square quantization error) solution, and the resulting pdf-optimized (non-uniform) quantizer is referred to as a *Lloyd–Max* quantizer, named after two people who independently developed iterative methods[8][21][22] to solve the two sets of simultaneous equations resulting from $\partial D / \partial b_k = 0$ and $\partial D / \partial y_k = 0$, as follows:

$$\frac{\partial D}{\partial b_k} = 0 \Rightarrow b_k = \frac{y_k + y_{k+1}}{2},$$

which places each threshold at the midpoint between each pair of reconstruction values, and

$$\frac{\partial D}{\partial y_k} = 0 \Rightarrow y_k = \frac{\int_{b_{k-1}}^{b_k} x f(x)dx}{\int_{b_{k-1}}^{b_k} f(x)dx} = \frac{1}{p_k} \int_{b_{k-1}}^{b_k} x f(x)dx$$

which places each reconstruction value at the centroid (conditional expected value) of its associated classification interval.

Lloyd's Method I algorithm, originally described in 1957, can be generalized in a straightforward way for application to vector data. This generalization results in the Linde–Buzo–Gray (LBG) or k-means classifier optimization methods. Moreover, the technique can be further generalized in a straightforward way to also include an entropy constraint for vector data.[23]

Uniform quantization and the 6 dB/bit approximation

The Lloyd–Max quantizer is actually a uniform quantizer when the input pdf is uniformly distributed over the range $[y_1 - \Delta/2,\ y_M + \Delta/2)$. However, for a source that does not have a uniform distribution, the minimum-distortion quantizer may not be a uniform quantizer.

The analysis of a uniform quantizer applied to a uniformly distributed source can be summarized in what follows:

A symmetric source X can be modelled with $f(x) = \frac{1}{2X_{max}}$, for $x \in [-X_{max}, X_{max}]$ and 0 elsewhere. The step size $\Delta = \frac{2X_{max}}{M}$ and the *signal to quantization noise ratio* (SQNR) of the quantizer is

$$\mathrm{SQNR} = 10\log_{10}\frac{\sigma_x^2}{\sigma_q^2} = 10\log_{10}\frac{(M\Delta)^2/12}{\Delta^2/12} = 10\log_{10}M^2 = 20\log_{10}M.$$

For a fixed-length code using N bits, $M = 2^N$, resulting in $\mathrm{SQNR} = 20\log_{10}2^N = N\cdot(20\log_{10}2) = N\cdot 6.0206\,\mathrm{dB}$,

or approximately 6 dB per bit. For example, for N=8 bits, M=256 levels and SQNR = 8*6 = 48 dB; and for N=16 bits, M=65536 and SQNR = 16*6 = 96 dB. The property of 6 dB improvement in SQNR for each extra bit used in quantization is a well-known figure of merit. However, it must be used with care: this derivation is only for a uniform quantizer applied to a uniform source.

For other source pdfs and other quantizer designs, the SQNR may be somewhat different from that predicted by 6 dB/bit, depending on the type of pdf, the type of source, the type of quantizer, and the bit rate range of operation.

However, it is common to assume that for many sources, the slope of a quantizer SQNR function can be approximated as 6 dB/bit when operating at a sufficiently high bit rate. At asymptotically high bit rates, cutting the step size in half increases the bit rate by approximately 1 bit per sample (because 1 bit is needed to indicate whether the value is in the left or right half of the prior double-sized interval) and reduces the mean squared error by a factor of 4 (i.e., 6 dB) based on the $\Delta^2/12$ approximation.

At asymptotically high bit rates, the 6 dB/bit approximation is supported for many source pdfs by rigorous theoretical analysis.[4][5][7][8] Moreover, the structure of the optimal scalar quantizer (in the rate–distortion sense) approaches that of a uniform quantizer under these conditions.[7][8]

Other fields

Many physical quantities are actually quantized by physical entities. Examples of fields where this limitation applies include electronics (due to electrons), optics (due to photons), biology (due to DNA), physics (due to Planck limits) and chemistry (due to molecules). This limitation is sometimes known in these fields as the "quantum noise limit".

See also

- Analog-to-digital converter
- Beta encoder
- Data binning
- Discretization

- Discretization error
- Posterization
- Pulse code modulation
- Quantile
- Regression dilution - a bias in parameter estimates caused by errors such as quantization in the explanatory or independent variable

Notes

1. Allen Gersho and Robert M. Gray, *Vector Quantization and Signal Compression (https://books.google.com/books/about/Vector_Quantization_and_Signal_Compressi.html?id=DwcDm6xgItUC)*, Springer, ISBN 978-0-7923-9181-4, 1991.
2. Hodgson, Jay (2010). *Understanding Records*, p.56. ISBN 978-1-4411-5607-5. Adapted from Franz, David (2004). *Recording and Producing in the Home Studio*, p.38-9. Berklee Press.
3. William Fleetwood Sheppard, "On the Calculation of the Most Probable Values of Frequency Constants for data arranged according to Equidistant Divisions of a Scale", *Proceedings of the London Mathematical Society*, Vol. 29, pp. 353–80, 1898.doi:10.1112/plms/s1-29.1.353 (https://doi.org/10.1112%2Fplms%2Fs1-29.1.353)
4. W. R. Bennett, "Spectra of Quantized Signals (http://www.alcatel-lucent.com/bstj/vol27-1948/articles/bstj27-3-446.pdf)", *Bell System Technical Journal*, Vol. 27, pp. 446–472, July 1948.
5. B. M. Oliver, J. R. Pierce, and Claude E. Shannon, "The Philosophy of PCM", *Proceedings of the IRE*, Vol. 36, pp. 1324–1331, Nov. 1948. doi:10.1109/JRPROC.1948.231941 (https://doi.org/10.1109%2FJRPROC.1948.231941)
6. Seymour Stein and J. Jay Jones, *Modern Communication Principles (https://books.google.com/books/about/Modern_communication_principles.html?id=jBc3AQAAIAAJ)*, McGraw–Hill, ISBN 978-0-07-061003-3, 1967 (p. 196).
7. Herbert Gish and John N. Pierce, "Asymptotically Efficient Quantizing", *IEEE Transactions on Information Theory*, Vol. IT-14, No. 5, pp. 676–683, Sept. 1968. doi:10.1109/TIT.1968.1054193 (https://doi.org/10.1109%2FTIT.1968.1054193)
8. Robert M. Gray and David L. Neuhoff, "Quantization", *IEEE Transactions on Information Theory*, Vol. IT-44, No. 6, pp. 2325–2383, Oct. 1998. doi:10.1109/18.720541 (https://doi.org/10.1109%2F18.720541)
9. Allen Gersho, "Quantization", *IEEE Communications Society Magazine*, pp. 16–28, Sept. 1977. doi:10.1109/MCOM.1977.1089500 (https://doi.org/10.1109%2FMCOM.1977.1089500)
10. Rabbani, Majid; Joshi, Rajan L.; Jones, Paul W. (2009). "Section 1.2.3: Quantization, in Chapter 1: JPEG 2000 Core Coding System (Part 1)". In Schelkens, Peter; Skodras, Athanassios; Ebrahimi, Touradj. *The JPEG 2000 Suite*. John Wiley & Sons. pp. 22–24. ISBN 978-0-470-72147-6.
11. Taubman, David S.; Marcellin, Michael W. (2002). "Chapter 3: Quantization". *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers. p. 107. ISBN 0-7923-7519-X.
12. Gary J. Sullivan, "Efficient Scalar Quantization of Exponential and Laplacian Random Variables", *IEEE Transactions on Information Theory*, Vol. IT-42, No. 5, pp. 1365–1374, Sept. 1996. doi:10.1109/18.532878 (https://doi.org/10.1109%2F18.532878)
13. Bernard Widrow, "A study of rough amplitude quantization by means of Nyquist sampling theory", *IRE Trans. Circuit Theory*, Vol. CT-3, pp. 266–276, 1956. doi:10.1109/TCT.1956.1086334 (https://doi.org/10.1109%2FTCT.1956.1086334)
14. Bernard Widrow, "Statistical analysis of amplitude quantized sampled data systems (http://www-isl.stanford.edu/~widrow/papers/j1961statisticalanalysis.pdf)", *Trans. AIEE Pt. II: Appl. Ind.*, Vol. 79, pp. 555–568, Jan. 1961.
15. Daniel Marco and David L. Neuhoff, "The Validity of the Additive Noise Model for Uniform Scalar Quantizers", *IEEE Transactions on Information Theory*, Vol. IT-51, No. 5, pp. 1739–1755, May 2005. doi:10.1109/TIT.2005.846397 (https://doi.org/10.1109%2FTIT.2005.846397)
16. Pohlman, Ken C. (1989). *Principles of Digital Audio 2nd Edition* (https://books.google.com/books?id=VZw6z9a03ikC&pg=PA37&source=gbs_selected_pages&cad=0_1). SAMS. p. 60.

17. Okelloto, Tom (2001). *The Art of Digital Audio 3rd Edition*. Focal Press. ISBN 0-240-51587-0.
18. Nariman Farvardin and James W. Modestino, "Optimum Quantizer Performance for a Class of Non-Gaussian Memoryless Sources", *IEEE Transactions on Information Theory*, Vol. IT-30, No. 3, pp. 485–497, May 1982 (Section VI.C and Appendix B). doi:10.1109/TIT.1984.1056920 (https://doi.org/10.1109%2FTIT.1984.1056920)
19. Toby Berger, "Optimum Quantizers and Permutation Codes", *IEEE Transactions on Information Theory*, Vol. IT-18, No. 6, pp. 759–765, Nov. 1972. doi:10.1109/TIT.1972.1054906 (https://doi.org/10.1109%2FTIT.1972.1054906)
20. Toby Berger, "Minimum Entropy Quantizers and Permutation Codes", *IEEE Transactions on Information Theory*, Vol. IT-28, No. 2, pp. 149–157, Mar. 1982. doi:10.1109/TIT.1982.1056456 (https://doi.org/10.1109%2FTIT.1982.1056456)
21. Stuart P. Lloyd, "Least Squares Quantization in PCM", *IEEE Transactions on Information Theory*, Vol. IT-28, pp. 129–137, No. 2, March 1982 doi:10.1109/TIT.1982.1056489 (https://doi.org/10.1109%2FTIT.1982.1056489) (work documented in a manuscript circulated for comments at Bell Laboratories with a department log date of 31 July 1957 and also presented at the 1957 meeting of the Institute of Mathematical Statistics, although not formally published until 1982).
22. Joel Max, "Quantizing for Minimum Distortion", *IRE Transactions on Information Theory*, Vol. IT-6, pp. 7–12, March 1960. doi:10.1109/TIT.1960.1057548 (https://doi.org/10.1109%2FTIT.1960.1057548)
23. Philip A. Chou, Tom Lookabaugh, and Robert M. Gray, "Entropy-Constrained Vector Quantization", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-37, No. 1, Jan. 1989. doi:10.1109/29.17498 (https://doi.org/10.1109%2F29.17498)

References

- Sayood, Khalid (2005), *Introduction to Data Compression, Third Edition*, Morgan Kaufmann, ISBN 978-0-12-620862-7
- Jayant, Nikil S.; Noll, Peter (1984), *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice–Hall, ISBN 978-0-13-211913-9
- Gregg, W. David (1977), *Analog & Digital Communication*, John Wiley, ISBN 978-0-471-32661-8
- Stein, Seymour; Jones, J. Jay (1967), *Modern Communication Principles*, McGraw–Hill, ISBN 978-0-07-061003-3

External links

- Quantization noise in Digital Computation, Signal Processing, and Control (http://www.mit.bme.hu/books/quantization/), Bernard Widrow and István Kollár, 2007.
- The Relationship of Dynamic Range to Data Word Size in Digital Audio Processing (https://web.archive.org/web/20060522134626/http://www.techonline.com/community/related_content/20771)
- Round-Off Error Variance (http://ccrma.stanford.edu/~jos/mdft/Round_Off_Error_Variance.html) — derivation of noise power of q²/12 for round-off error
- Dynamic Evaluation of High-Speed, High Resolution D/A Converters (http://www.ieee.li/pdf/essay/dynamic_evaluation_dac.pdf) Outlines HD, IMD and NPR measurements, also includes a derivation of quantization noise
- Signal to quantization noise in quantized sinusoidal (http://www.dsplog.com/2007/03/19/signal-to-quantization-noise-in-quantized-sinusoidal/)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Quantization_(signal_processing)&oldid=802951553"

- This page was last edited on 29 September 2017, at 14:40.
-