7.6 Quantization Noise

For most of the engineering applications the input signal is continuous in time or analog wave form. This signal is to be converted into digital by using ADC. The process of converting an analog signal to a digital is shown in Fig. 7.1. At first the signal $x(t)$ is sampled at regular intervals $t = nT$ where $n = 0, 1, 2 \ldots$ to create a sequence $x(n)$. This is done by a sampler. Then numeric equivalent of each sample $x(n)$ is expressed by a finite number of bits giving the sequence $x_q(n)$. The difference signal $e(n) = x_q(n) - x(n)$ is called quantization noise or A/D conversion noise.

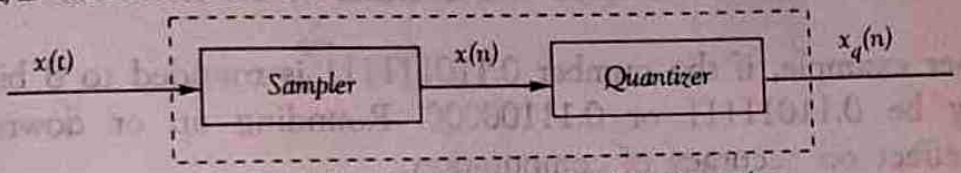

Fig. 7.1 Block diagram of A/D converter

Let us assume a sinusoidal signal varying between $+1$ and -1 having a dynamic range 2. If the ADC used to convert the sinusoidal signal employs $(b+1)$ bits including sign bit, the number of levels available for quantizing $x(n)$ is 2^{b+1}. Thus the interval between successive levels is

$$q = \frac{2}{2^{b+1}} = 2^{-b} \qquad \ldots (7.3)$$

where q is known as quantization step size.

If $b = 3$ bits then $q = 2^{-3} = 0.125$

The common methods of quantization are

1. Truncation.
2. Rounding

7.6.1 Truncation

Truncation is a process of discarding all bits less significant than least significant bit that is retained.

Suppose, if we truncate the following binary number from 8 bits to 4 bits we obtain

$$0.00110011 \text{ to } 0.0011$$
$$\underset{8 \text{ bits}}{} \qquad \underset{4 \text{ bits}}{}$$

$$1.01001001 \text{ to } 1.0100$$
$$\underset{8 \text{ bits}}{} \qquad \underset{4 \text{ bits}}{}$$

When we truncate the number, the signal value is approximated by the highest qauntization level that is not greater than the signal.

7.6.2 Rounding

Rounding of a number of b bits is accomplished by choosing the rounded result as the b bit number closest to the original number unrounded. For example, 0.11010 rounded to three bits is either 0.110 or 0.111.

Another example, if the number 0.110111111 is rounded to 8 bits then the result may be 0.11011111 or 0.11100000. Rounding up or down will have negligible effect on accuracy of computation.

7.6.3 Error due to truncation and rounding

If the quantization method is truncation, the number is approximated by the nearest level that does not exceed it. In this case the error $x_T - x$ is negative or zero where x_T is truncated value of x and it is assumed $|x| \leq 0$

The error made by truncating a number to b bits following the binary point satisfies the inequality,

$$0 \geq x_T - x > -2^{-b} \qquad \ldots (7.4)$$

For example, consider the decimal number 0.12890625. It's binary equivalent is 0.00100001. If we truncate the binary number to 4 bits, we have $x_T = (0.0010)_2$ whose decimal value is 0.125. Now the error $(x_T - x) = -0.00390625$, which is greater than $-2^{-b} = -2^{-4} = -0.0625$ satisfying the inequality given in Eq. (7.4).

The Eq. (7.4) holds for both sign-magnitude, one's-complement and two's-complement if $x > 0$. If $x < 0$, we have to find whether the Eq. (7.4) holds good for all types of representations.

Consider first the two's complement representation. From Eq. (7.2) the magnitude of the negative number is

$$x = 1 - \sum_{i=1}^{b} c_i 2^{-i}$$

If we truncate the number to N bits then

$$x_T = 1 - \sum_{i=1}^{N} c_i 2^{-i}$$

The change in magnitude

$$x_T - x = \sum_{i=1}^{b} c_i 2^{-i} - \sum_{i=1}^{N} c_i 2^{-i}$$

$$= \sum_{i=N}^{b} c_i 2^{-i}$$

$$= \geq 0 \qquad \ldots (7.5a)$$

From the Eq. (7.5a) we find, the truncation increases the magnitude, which implies that error is negative and satisfy the inequality

$$0 \geq x_T - x \geq -2^{-b} \qquad \ldots (7.5b)$$

For one's complement representation the magnitude of negative number with b bits is given by

$$x = 1 - \sum_{i=1}^{b} c_i 2^{-i} - 2^{-b} \qquad \ldots (7.6a)$$

when the number is truncated to N bits, then

$$x_T = 1 - \sum_{i=1}^{N} c_i 2^{-i} - 2^{-N} \qquad \ldots (7.6b)$$

The change in magnitude due to truncation is

$$x_T - x = \sum_{i=N}^{b} c_i 2^{-i} - (2^{-N} - 2^{-b})$$

$$< 0 \qquad \ldots (7.7)$$

Therefore the magnitude decreases with truncation which implies that error is positive and satisfy the inequality

$$0 \le x_T - x < 2^{-b} \qquad \ldots (7.8)$$

The Eq. (7.8) holds good for sign magnitude representation also.

In floating point systems the effect of truncation is visible only in the mantissa. Let the mantissa is truncated to N bits.

If $x = 2^c \cdot M$, then

$$x_T = 2^c \cdot M_T \qquad \ldots (7.9)$$

Error $e = x_T - x = 2^c (M_T - M)$ ... (7.10)

From Eq. (7.5b), with two's complement representation of mantissa, we have

$$0 \ge M_T - M > -2^{-b} \qquad \ldots (7.11)$$

$$0 \ge e > -2^{-b} 2^c \qquad \ldots (7.12)$$

We define relative error $\varepsilon = \dfrac{x_T - x}{x} = \dfrac{e}{x}$ ... (7.13)

Now Eq. (7.12) can be written as

$$0 \ge \varepsilon x > -2^{-b} 2^c \qquad \ldots (7.14)$$

or $0 \ge \varepsilon 2^c M > -2^{-b} 2^c$ ... (7.15)

or $0 \ge \varepsilon M > -2^{-b}$ ... (7.16)

if $M = 1/2$, the relative error is maximum

Therefore,

$$0 \ge \varepsilon > -2 \cdot 2^{-b} \qquad \ldots (7.17)$$

if $M = -1/2$ the relative error range is

$$0 \le \varepsilon < 2 \cdot 2^{-b} \qquad \ldots (7.18)$$

In one's complement representation the error for truncation of positive values of the mantissa is

$$0 \geq M_T - M > -2^{-b} \quad \ldots (7.19)$$

or

$$0 \geq e > -2^{-b} \cdot 2^c \quad \ldots (7.20)$$

with $e = \varepsilon x = \varepsilon\, 2^c \cdot M$ ... (7.21)

and $M = 1/2$, we get the maximum range of the relative error for positive M as

$$0 \geq \varepsilon > -2 \cdot 2^{-b} \quad \ldots (7.22)$$

For negative mantissa values the error is

$$0 \leq M_T - M < 2^{-b} \quad \ldots (7.23)$$

or

$$0 \leq e < 2^c\, 2^{-b} \quad \ldots (7.24)$$

with $M = -1/2$. The maximum range of the relative error for negative M is

$$0 \geq \varepsilon \geq -2 \cdot 2^{-b} \quad \ldots (7.25)$$

which is the same as positive M (Eq. (7.22)).

The probability density function $P(e)$ for truncation of fixed point and floating point numbers are shown in Fig. 7.2.

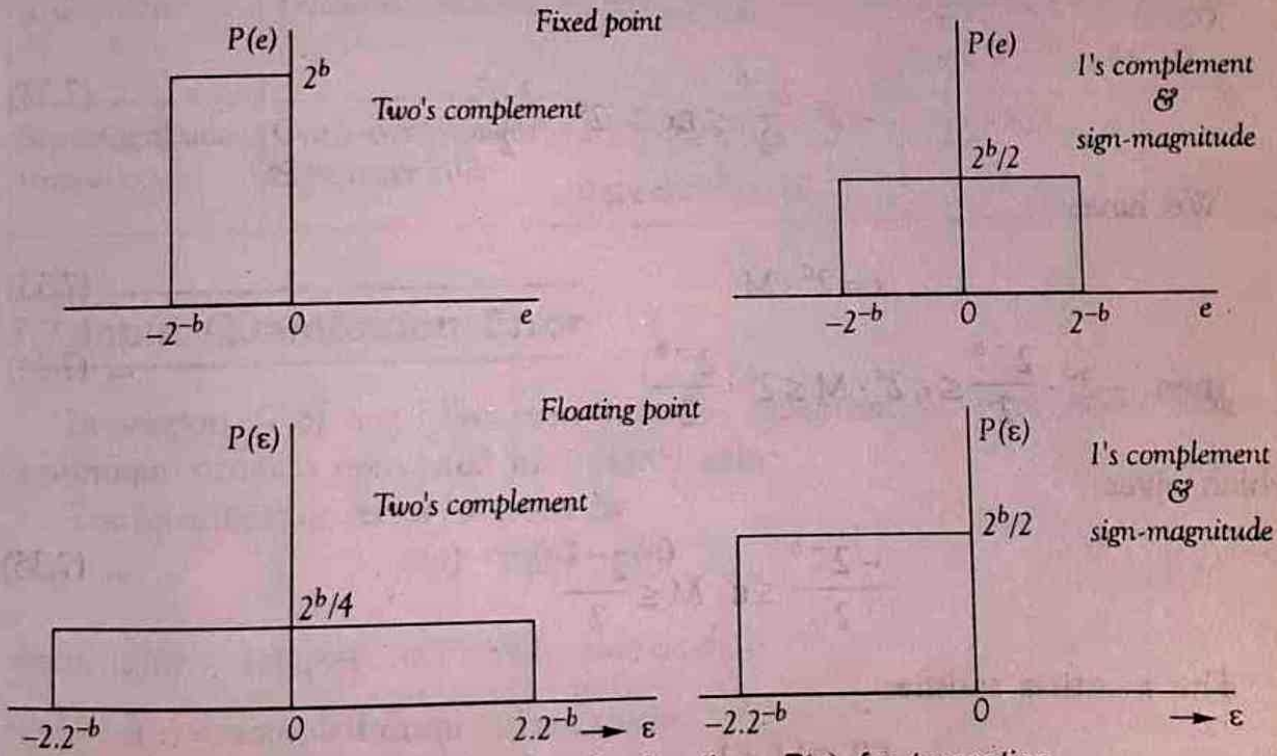

***Fig. 7.2** Probability density functions P(e) for truncation.*

In fixed point arithmetic the error due to rounding a number to b bits produces an error $e = x_T - x$ which satisfies the inequality,

$$\frac{-2^{-b}}{2} \le x_T - x \le \frac{2^{-b}}{2} \quad \ldots (7.26)$$

This is because with rounding, if the value lies half way between two levels, it can be approximated to either nearest higher level or by the nearest lower level. For fixed-point numbers Eq. (7.26) satisfies regardless of whether sign-magnitude, one's complement or two's-complement is used for negative numbers.

In floating-point arithmetic, only the mantissa is affected by quantization

if $x = M \cdot 2^c$ ... (7.27)

and $x_T = M_T \cdot 2^c$ ... (7.28)

Then $e = x_T - x = (M_T - M)\, 2^c$ ... (7.29)

But for rounding

$$\frac{-2^{-b}}{2} \le M_T - M \le \frac{2^{-b}}{2} \quad \ldots (7.30)$$

Using Eq. (7.29), the Eq. (7.30) can be written as

$$-2^c \frac{2^{-b}}{2} \le x_T - x \le 2^c \frac{2^{-b}}{2} \quad \ldots (7.31)$$

or

$$-2^c \frac{2^{-b}}{2} \le \varepsilon x \le 2^c \frac{2^{-b}}{2} \quad \ldots (7.32)$$

We have

$$x = 2^c \cdot M \quad \ldots (7.33)$$

then $-2^c \cdot \frac{2^{-b}}{2} \le \varepsilon\, 2^c \cdot M \le 2^c \cdot \frac{2^{-b}}{2}$ ... (7.34)

which gives

$$\frac{-2^{-b}}{2} \le \varepsilon \cdot M \le \frac{2^{-b}}{2} \quad \ldots (7.35)$$

The mantissa satisfies

$$1/2 \le M < 1$$

If $M = 1/2$ we get the maximum range of relative error

$$-2^{-b} \le \varepsilon < 2^{-b} \quad \ldots (7.36)$$

If $M = \frac{1}{2}$, we get the maximum range of relative error

$$-2^{-b} \le \varepsilon < 2^{-b} \qquad \ldots (7.36)$$

The probability density function for rounding is shown in Fig. 7.3.

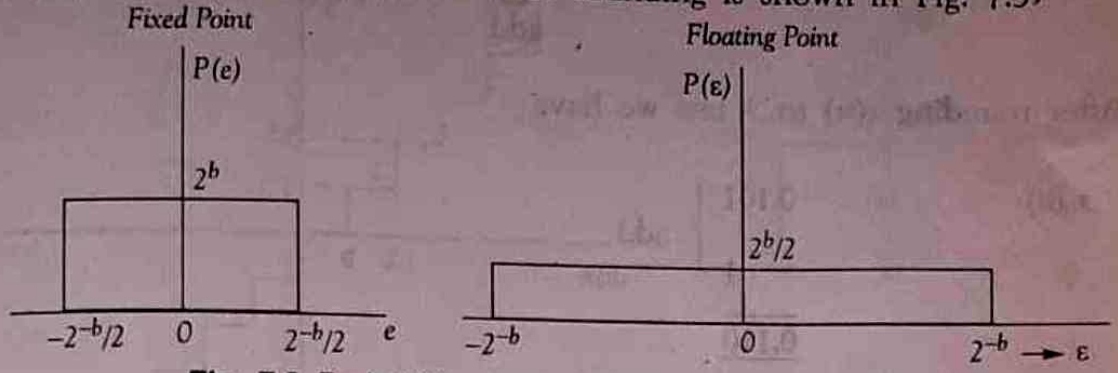

Fig. 7.3 Probability density function P(ε) for rounding

Table 7.2 Quantization error ranges

Type of quantization	Type of arithmetic	Fixed-Point number range	Floating-Point number Relative-Error range
Rounding	Sign-magnitude One's-complement Two's-complement	$\frac{-2^{-b}}{2} \le e \le \frac{2^b}{2}$	$-2^{-b} \le \varepsilon \le 2^{-b}$
Truncation	Two's-complement	$-2^{-b} < e \le 0$	$-2.2^{-b} < \varepsilon \le 0, M > 0$ $0 \le \varepsilon < 2.2^{-b}, M < 0$
Sign-magnitude truncation	One's-complement Sign-magnitude	$-2^{-b} < e \le 0, x > 0$ $0 \le e \le 2^{-b}, x < 0$	$-2.2^{-b} < \varepsilon \le 0$

7.7 Input Quantization Error

In section (7.6) we have seen that the quantization error arises when a continuous signal is converted into digital value.

The quantization error is given by

$$e(n) = x_q(n) - x(n) \qquad \ldots (7.37)$$

where $x_q(n)$ = sampled quantized value and

$x(n)$ = sampled unquantized value.

Depending on the way in which $x(n)$ is quantized different distributions of quantization noise may be obtained. If rounding of a number is used to get $x_q(n)$ then the error signal satisfies the relation

$$-\frac{q}{2} \le e(n) \le \frac{q}{2} \qquad \ldots (7.38)$$

because the quantized signal may be greater or less than actual signal.

For example, let $x(n) = (0.70)_{10} = (0.10110011 \ldots)_2$

add

After rounding $x(n)$ to 3 bits we have

$$\begin{aligned} x_q(n) &= 0.101 \\ &\quad\ \ 1 \quad \text{add} \\ &\quad\ \overline{0.110} \\ &= (0.75)_{10} \end{aligned}$$

Now the error

$$e(n) = x_q(n) - x(n) = 0.05$$

which satisfies the inequality.

The probability density function $P(e)$ for roundoff error and quantization characteristics with rounding is shown in Fig. 7.4a and Fig. 7.4b respectively.

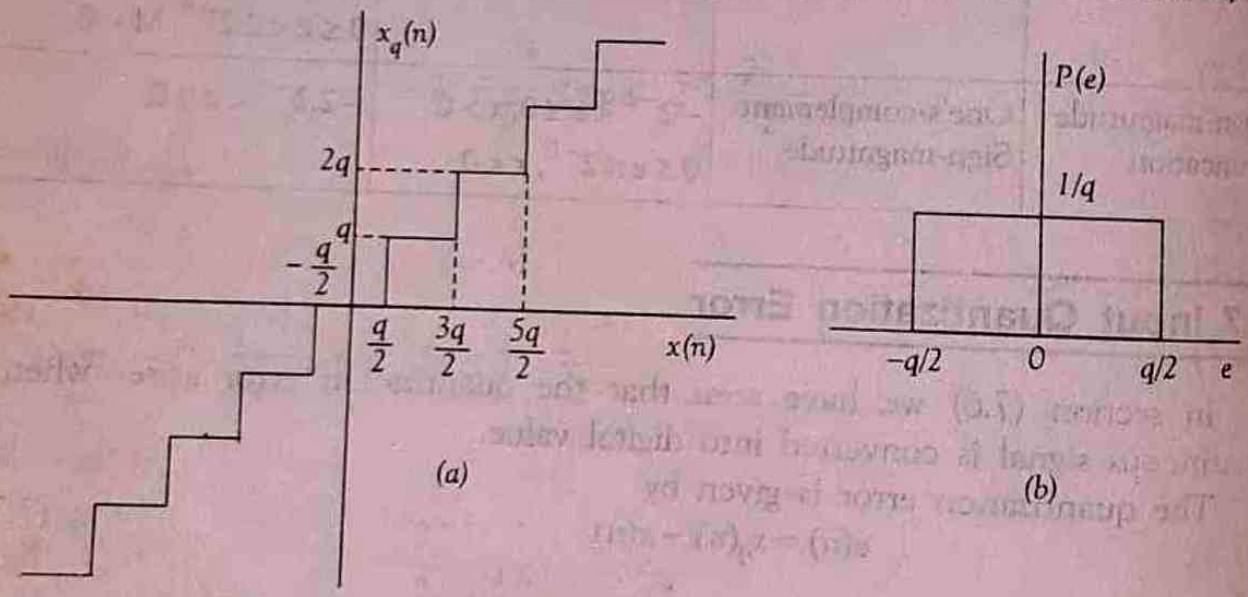

Fig. 7.4 (a) Quantizer characteristics with rounding
(b) Probability density function for roundoff error

The other type of quantization can be obtained by truncation. In truncation the signal is represented by the highest quantization level that is not greater than the signal. Therefore, in two's-complement truncation, the error $e(n)$ is always negative and satisfies the inequality.

$$-q \le e(n) < 0 \qquad \text{... (7.39)}$$

The quantizer characteristics for truncation and probability density function $P(e)$ for two's-complement truncation is shown in Fig. 7.5a and Fig. 7.5b respectively.

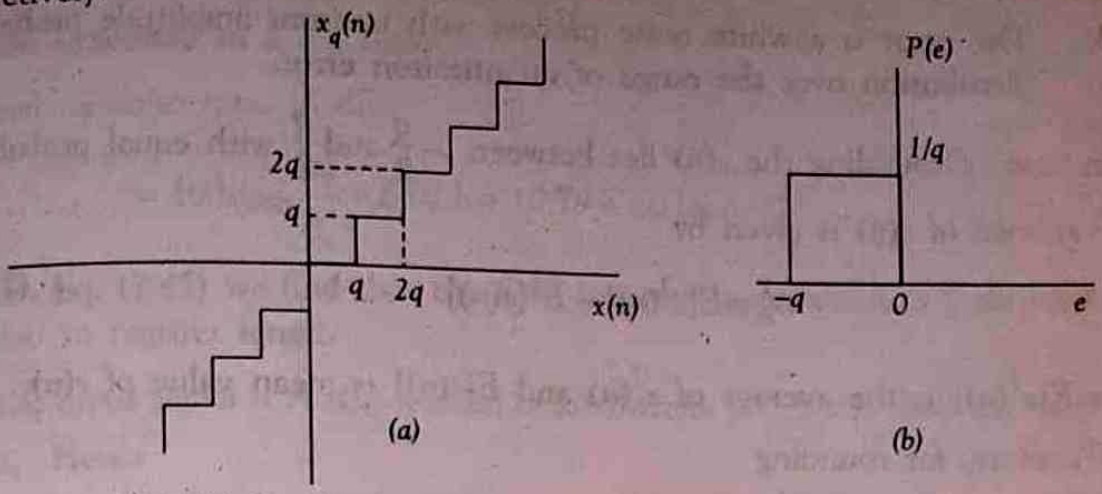

***Fig. 7.5** (a) Quantizer characteristic with two's-complement truncation (b) Probability density function of truncation error*

From the Fig. 7.4 and Fig. 7.5 it is clear that the quantization error mean value is 0 for rounding and $-q/2$ for two's-complement truncation.

7.7.1 Steady State Input Noise Power

In digital processing of analog signals, the quantization error is commonly viewed as an additive noise signal, that is

$$x_q(n) = x(n) + e(n) \qquad \text{... (7.40)}$$

Therefore, the A/D convertor output is the sum of the input signal $x(n)$ and the error signal $e(n)$ as shown in Fig. 7.6.

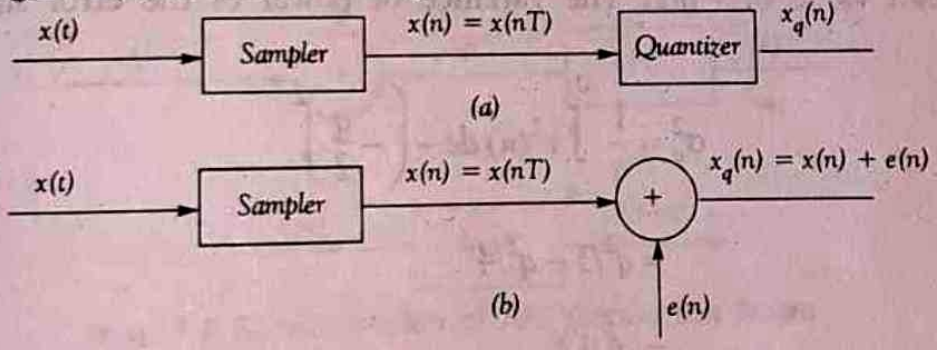

***Fig 7.6** Quantization noise model*

If the rounding is used for quantization then the quantization error $e(n) = x_q(n) - x(n)$ is bounded by

$$-q/2 \le e(n) \le q/2$$

In most cases, we can assume that the analog-to-digital conversion error $e(n)$ has the following properties.

1. The error sequence $e(n)$ is a sample sequence of a stationary random process.
2. The error sequence is uncorrelated with $x(n)$ and other signals in the system.
3. The error is a white noise process with uniform amplitude probability distribution over the range of quantizaiton error.

In case of rounding the $e(n)$ lies between $-\frac{q}{2}$ and $\frac{q}{2}$ with equal probability. The variance of $e(n)$ is given by

$$\sigma_e^2 = E[e^2(n)] - E^2[e(n)] \qquad \ldots (7.41a)$$

where $E[e^2(n)]$ is the average of $e^2(n)$ and $E[e(n)]$ is mean value of $e(n)$.

Therefore, for rounding

$$\sigma_e^2 = \frac{1}{q}\int_{-q/2}^{q/2} e^2(n)de - (0)^2 = \frac{q^2}{12}$$

Substituting Eq. (7.3) in Eq. (7.41b) we have

$$\sigma_e^2 = \frac{(2^{-b})^2}{12} = \frac{2^{-2b}}{12}$$

$$\boxed{\sigma_e^2 = \frac{2^{-2b}}{12}} \qquad \ldots (7.42)$$

In case of two's-complement truncation the $e(n)$ lies between 0 and $-q$ having mean value of $-q/2$. The variance or power of the error signal $e(n)$ is given by

$$\sigma_e^2 = \frac{1}{q}\int_{-q}^{0} e^2(n)\, de - \left(-\frac{q}{2}\right)^2$$

$$= q^2/3 - q^2/4$$

$$= q^2/12$$

Thus we have $\boxed{\sigma_e^2 = \frac{2^{-2b}}{12}}$ ... (7.43)

In both cases the value $\sigma_e^2 = \frac{2^{-2b}}{12}$, which is also known as the steady state noise power due to input quantization.

If the input signal is $x(n)$ and it's variance is σ_x^2, then the ratio of signal power to noise power which is known as signal to noise ratio for rounding is

$$\frac{\sigma_x^2}{\sigma_e^2} = \frac{\sigma_x^2}{2^{-2b}/12} = 12\,(2^{2b}\,\sigma_x^2) \qquad \ldots (7.44)$$

when expressed in a log scale

signal to noise ratio in dB

$$= 10 \log_{10} \frac{\sigma_x^2}{\sigma_e^2} = 6.02\,b + 10.79 + 10 \log_{10} \sigma_x^2 \qquad \ldots (7.45)$$

From Eq. (7.45) we find that the SNR increases approximately 6 dB for each bit added to register length.

If the input signal is $A\,x(n)$ instead of $x(n)$ where $0 < A < 1$, then the variance is $A^2\,\sigma_x^2$. Hence

$$SNR = 10 \log_{10} \frac{\sigma_x^2 A^2}{\sigma_e^2} = 6b + 10.8 + 10 \log_{10} \sigma_x^2 + 20 \log_{10} A \qquad \ldots (7.46)$$

$$\text{if } A = \frac{1}{4\,\sigma_x},\ SNR = 6\,b - 1.24\,dB \qquad \ldots (7.47)$$

Thus, to obtain $SNR \geq 80\,dB$ requires $b = 14\ bits$.

7.7.2 Steady State Output Noise Power

Due to A/D conversion noise one can represent the quantized input to a digital system with impulse response $h(n)$ as shown in Fig. 7.7.

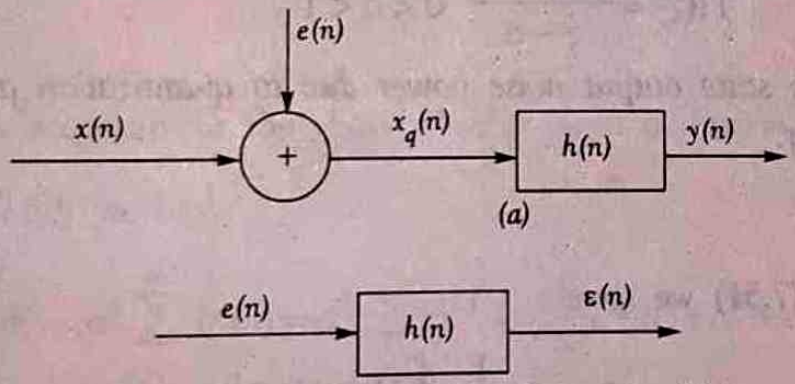

Fig. 7.7 Representation of A/D Conversion Noise

Let $\varepsilon(n)$ be the output noise due to quantization of the input. Then we get

$$\varepsilon(n) = e(n) * h(n)$$

$$= \sum_{k=0}^{n} h(k)\,e(n-k) \qquad \ldots (7.48)$$

(Refer section 8.11)

The variance of any term in the above sum is equal to $\sigma_e^2 h^2(n)$.

The variance of the sum of independent random variables is the sum of their variances. If the quantization errors are assumed to be independent at different sampling instances, then the variance of the output

$$\sigma_\varepsilon^2(n) = \sigma_e^2 \sum_{n=0}^{k} h^2(n) \quad \ldots (7.49)$$

To find the steady state variance, extend the limit k upto infinity.

Then, we have

$$\sigma_\varepsilon^2 = \sigma_e^2 \sum_{n=0}^{\infty} h^2(n) \quad \ldots (7.50)$$

Using Parseval's theorem the steady state output noise variance due to the quantization error is given by

$$\sigma_\varepsilon^2 = \sigma_e^2 \sum_{n=0}^{\infty} h^2(n) = \frac{\sigma_e^2}{2\pi j} \oint_C H(z)\, H(z^{-1})\, z^{-1}\, dz \quad \ldots (7.51)$$

(Refer section 8.11)

where the closed contour of integration is around the unit circle $|z| = 1$ in which case only the poles that lie inside the unit circle are evaluated using the residue theorem.

Example 7.3: *The output signal of an A/D converter is passed through a first order lowpass filter, with transfer function given by*

$$H(z) = \frac{(1-a)\,z}{z-a} \quad 0 < a < 1$$

find the steady state output noise power due to quantization at the output of the digital filter.

Solution

From Eq. (7.51) we have

$$\sigma_\varepsilon^2 = \sigma_e^2 \frac{1}{2\pi j} \oint_C H(z)\, H(z^{-1})\, z^{-1} dz$$

$$\text{Given } H(z) = \frac{(1-a)\,z}{(z-a)}$$

$$\text{then} \quad H(z^{-1}) = \frac{(1-a)\,z^{-1}}{(z^{-1}-a)}$$

Substituting $H(z)$ and $H(z^{-1})$ in Eq. (7.51), we have

$$\sigma_\varepsilon^2 = \sigma_e^2 \frac{1}{2\pi j} \oint_C \frac{(1-a)^2 z^{-1} dz}{(z-a)(z^{-1}-a)}$$

$$= \sigma_e^2 \left[\text{residue of } H(z) H(z^{-1}) z^{-1} \text{ at } z = a \right.$$

$$\left. + \text{residue of } H(z) H(z^{-1}) z^{-1} \text{ at } z = \frac{1}{a} \right]$$

$$= \sigma_e^2 \left[\cancel{(z-a)} \frac{(1-a)^2 z^{-1}}{\cancel{(z-a)}(z^{-1}-a)} \Bigg|_{z=a} + 0 \right]$$

residue at $z = 1/a$ is equal to zero as $\frac{1}{a} > 1$

$$= \sigma_e^2 \left[\frac{(1-a)^2}{1-a^2} \right] = \sigma_e^2 \left[\frac{1-a}{1+a} \right]$$

where $\sigma_e^2 = \frac{2^{-2b}}{12}$

Example 7.4: *Find the steady state variance of the noise in the output due to quantization of input for the first order filter.*

$$y(n) = ay(n-1) + x(n)$$

Solution

The impulse response for the above filter is given by $h(n) = a^n u(n)$

From Eq. (7.50) we have

$$\sigma_\varepsilon^2 = \sigma_e^2 \sum_{k=0}^{\infty} h^2(n) = \sigma_e^2 \sum_{k=0}^{\infty} a^{2n} = \sigma_e^2 \left[1 + a^2 + a^4 + \ldots \infty \right]$$

$$= \sigma_e^2 \frac{1}{1-a^2}$$

$$= \frac{2^{-2b}}{12} \left[\frac{1}{1-a^2} \right]$$

or

Given $y(n) = ay(n-1) + x(n)$

Taking z-transform on both sides we have

$$Y(z) = az^{-1} Y(z) + X(z)$$

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{1 - az^{-1}} = \frac{z}{z - a}$$

$$H(z^{-1}) = \frac{z^{-1}}{z^{-1} - a}$$

We know

$$\sigma_\varepsilon^2 = \sigma_e^2 \frac{1}{2\pi j} \oint_C H(z) H(z^{-1}) z^{-1} dz$$

Substituting $H(z)$ and $H(z^{-1})$ values in the above equation we get

$$\sigma_\varepsilon^2 = \sigma_e^2 \frac{1}{2\pi j} \oint_C H(z) H(z^{-1}) z^{-1} dz$$

$$\sigma_\varepsilon^2 = \sigma_e^2 \frac{1}{2\pi j} \oint_C \frac{z}{z-a} \frac{z^{-1}}{z^{-1} - a} z^{-1} dz$$

$$\sigma_\varepsilon^2 = \sigma_e^2 \frac{1}{2\pi j} \oint \frac{z^{-1}}{(z-a)(z^{-1} - a)} dz$$

$$= \sigma_\varepsilon^2 \left[\text{residue of } \frac{z^{-1}}{(z-a)(z^{-1} - a)} \text{ at } z = a \right.$$

$$\left. + \underbrace{\text{residue of } \frac{z^{-1}}{(z-a)(z^{-1} - a)} \text{ at } z = 1/a}_{\text{equal to zero}} \right]$$

$$= \sigma_e^2 \left[\cancel{(z-a)} \frac{z^{-1}}{\cancel{(z-a)}(z^{-1} - a)} \Bigg|_{z=a} \right]$$

$$= \sigma_e^2 \frac{a^{-1}}{a^{-1} - a} = \sigma_e^2 \frac{1}{1 - a^2}$$