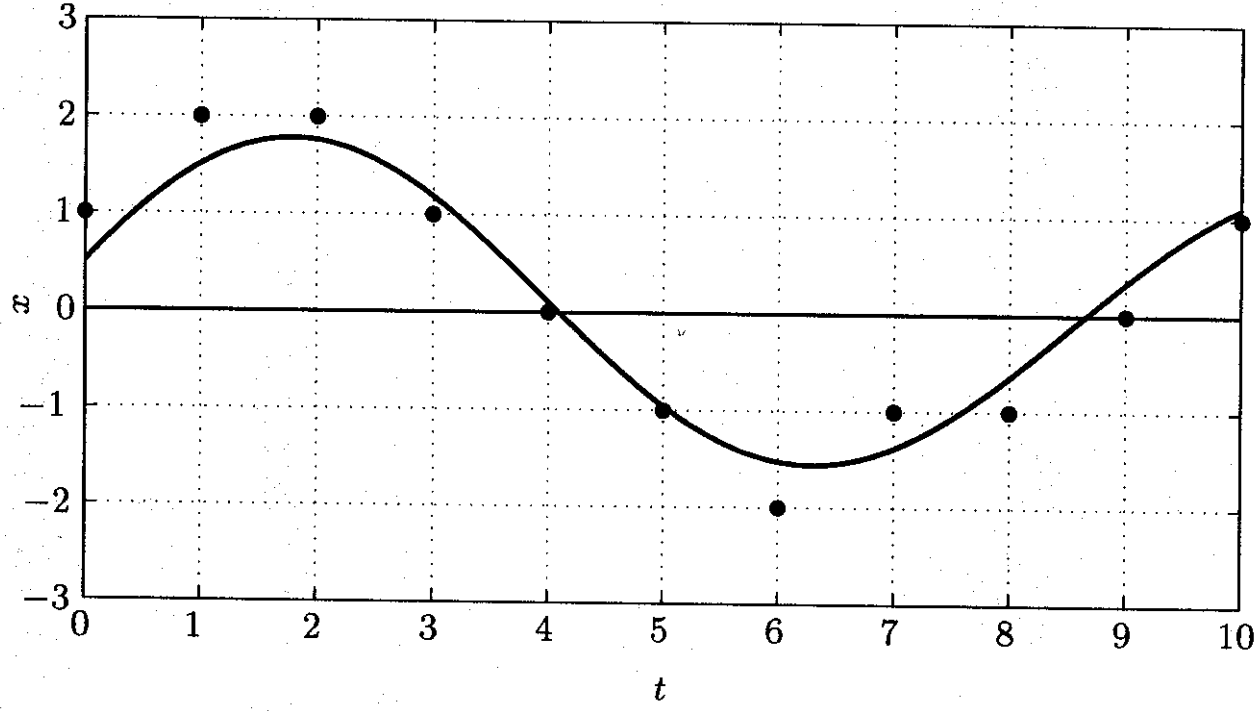
Quantization Noise

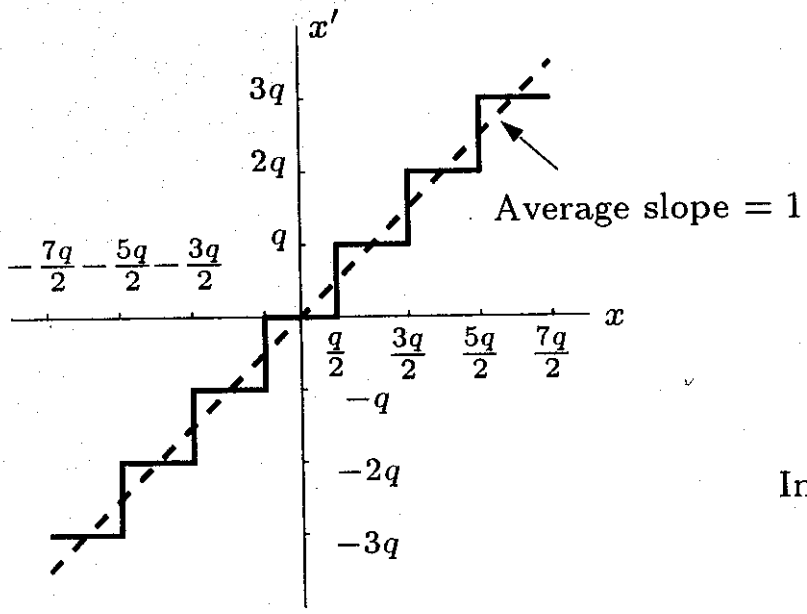
Bernard Widrow, Istvan Kollar, and Ming-Chang Liu

Information Systems Laboratory
Department of Electrical Engineering
Stanford University

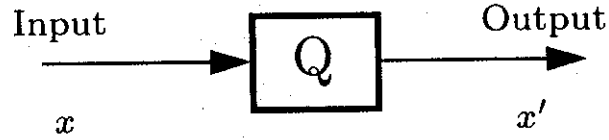
Sampling and Quantization



A Basic Quantizer

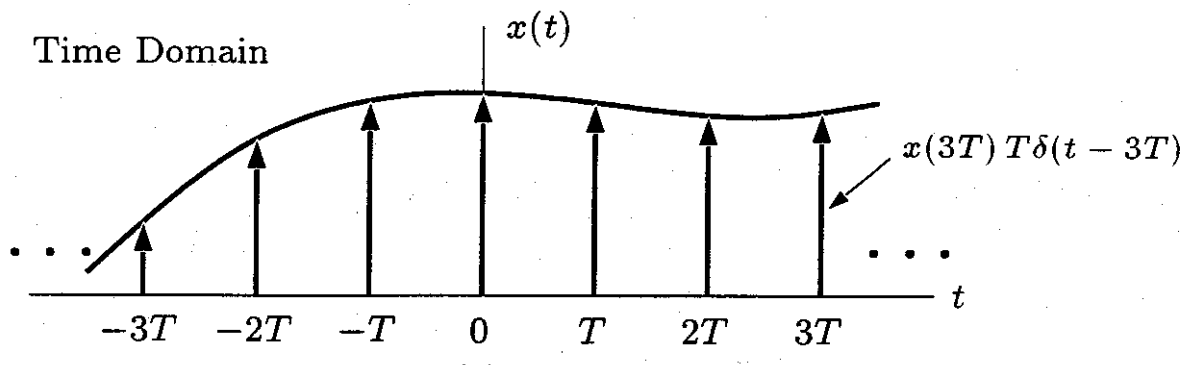


(a)

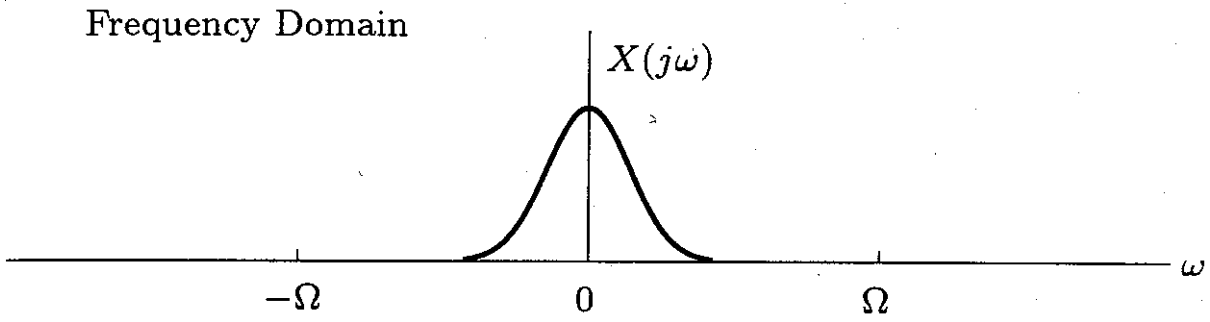


(b)

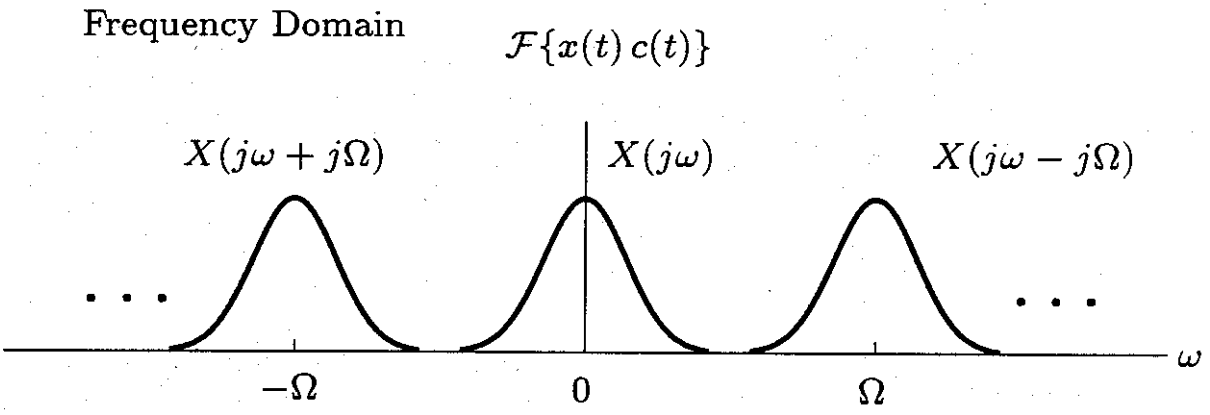
Fourier Transform of a Function and its Samples



(a)

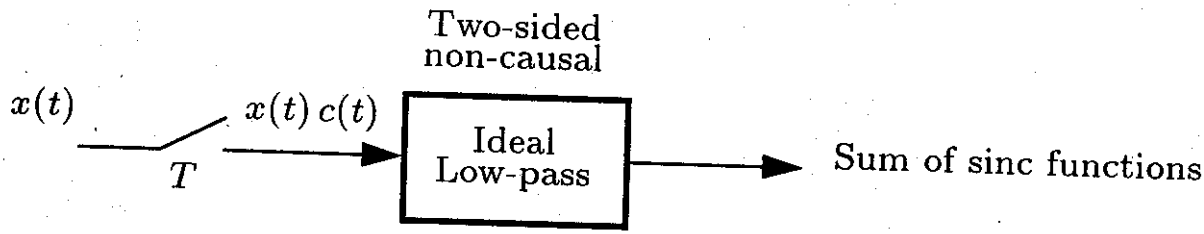


(b)



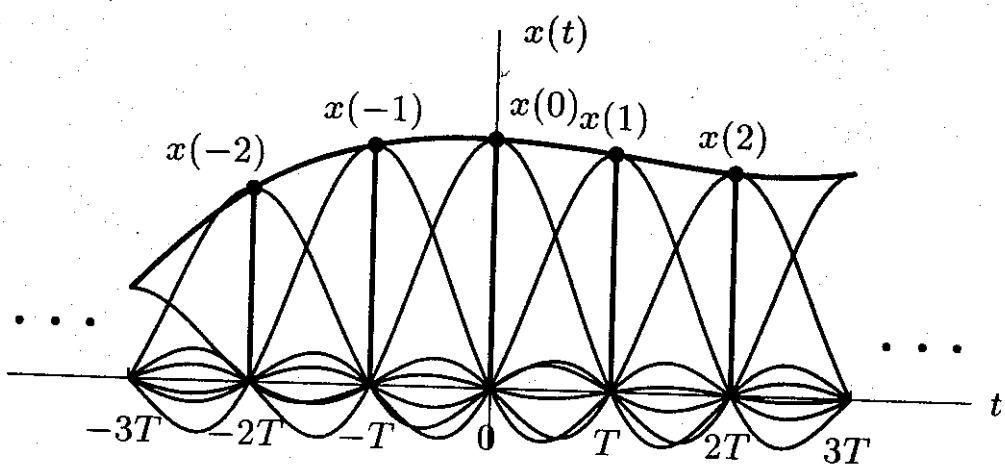
(c)

Recovery of Original Signal from its Samples



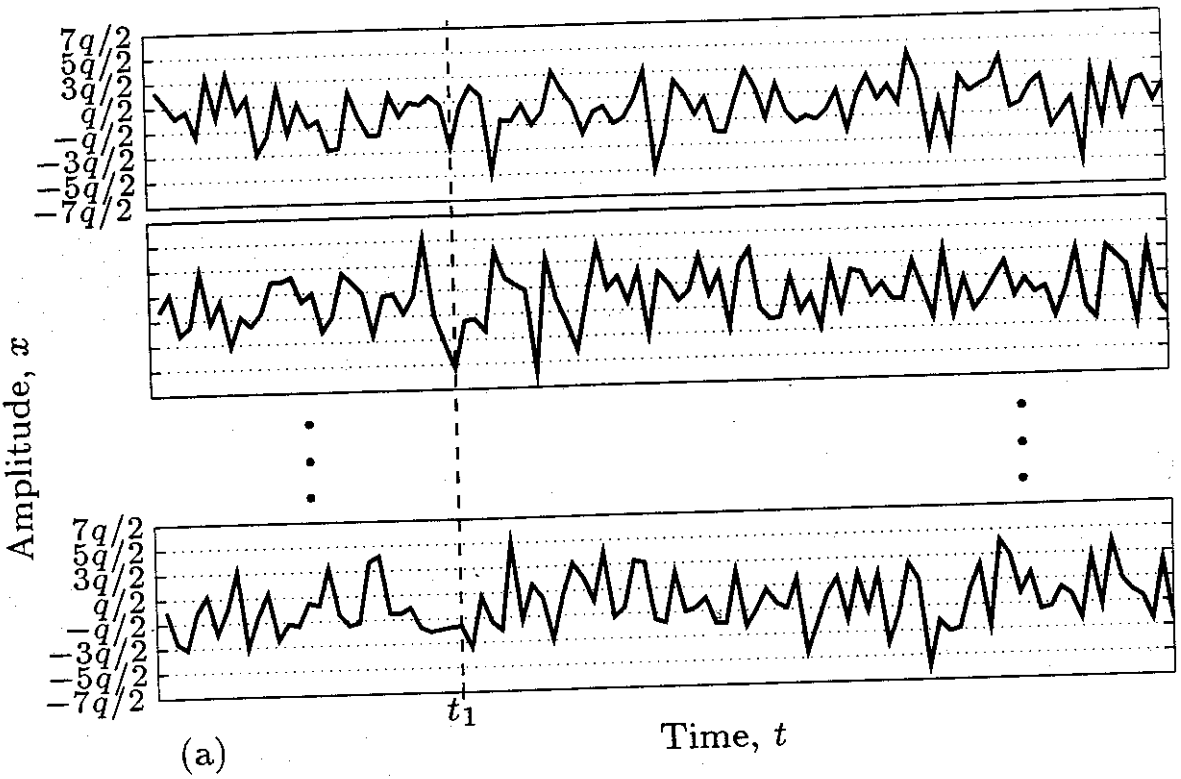
$$\frac{1}{T}\,\mathrm{sinc}\left(\frac{\pi t}{T}\right) = \frac{1}{T}\,\frac{\sin\left(\frac{\pi t}{T}\right)}{\frac{\pi t}{T}}$$

(a)

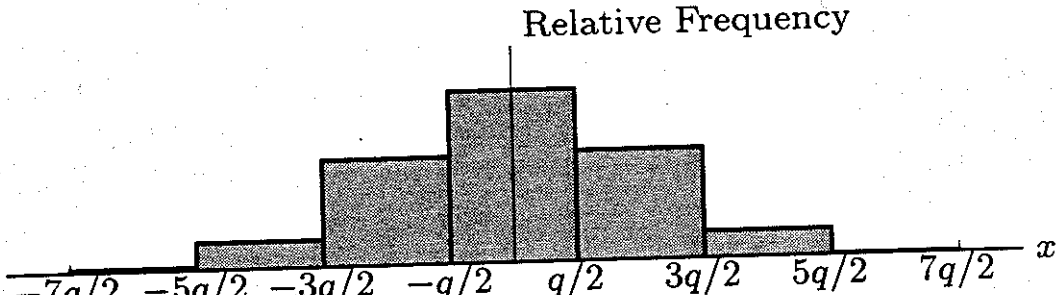


(b)

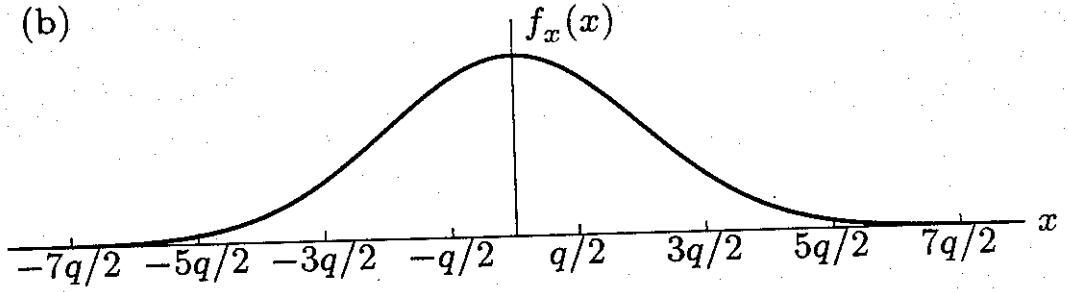
Derivation of a Histogram



(a)

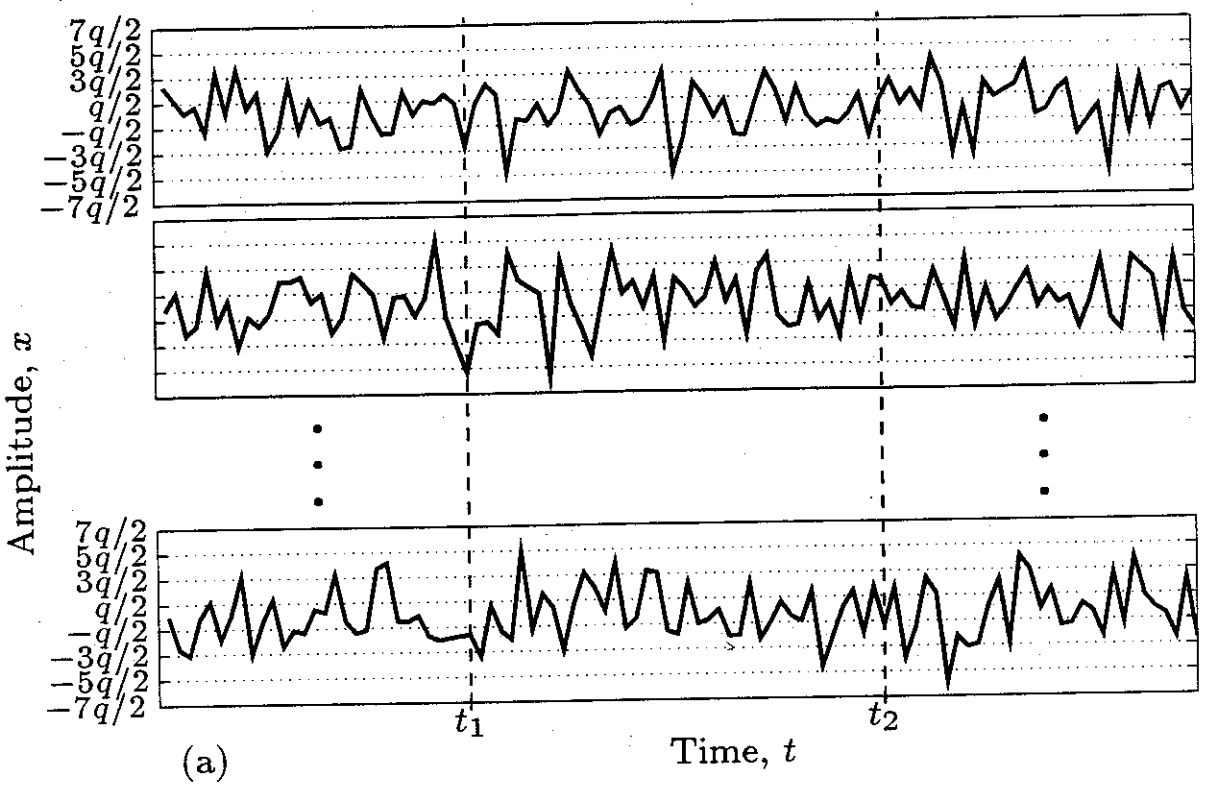


(b)

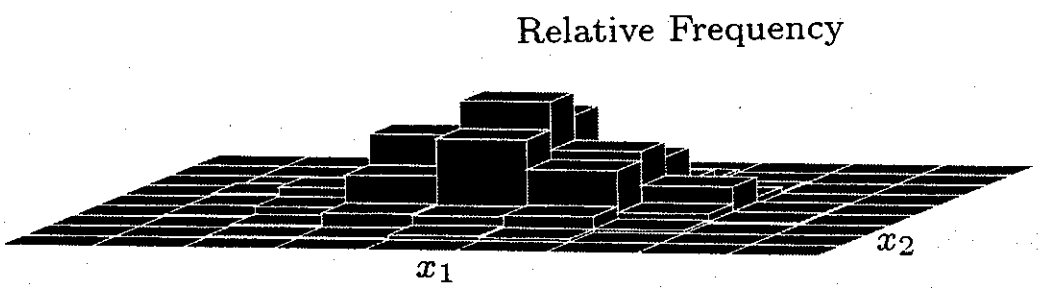


(c)

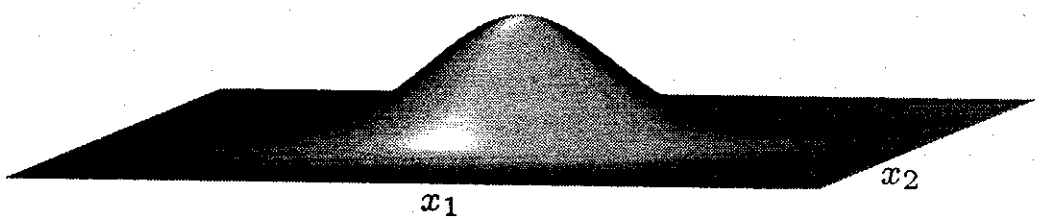
Derivation of a Two-Dimensional Histogram



(a)

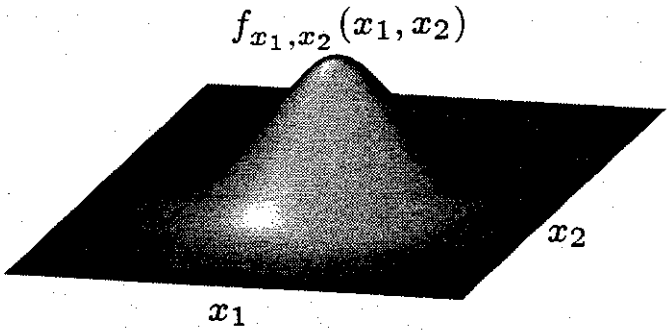


(b)

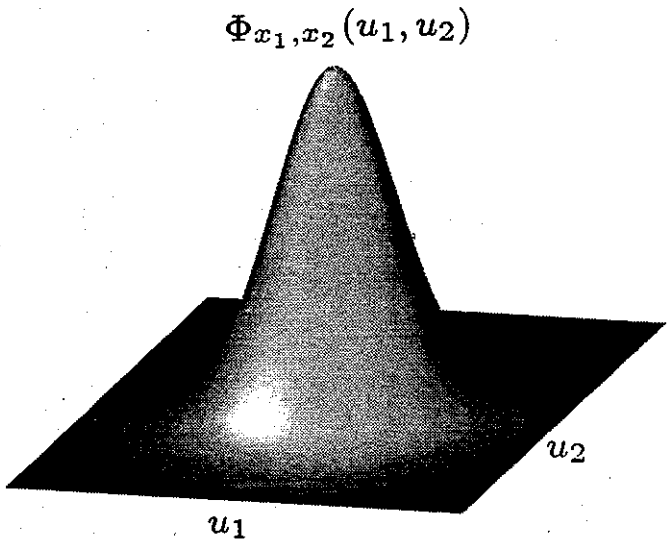


(c)

Pdf and Cf



(a)

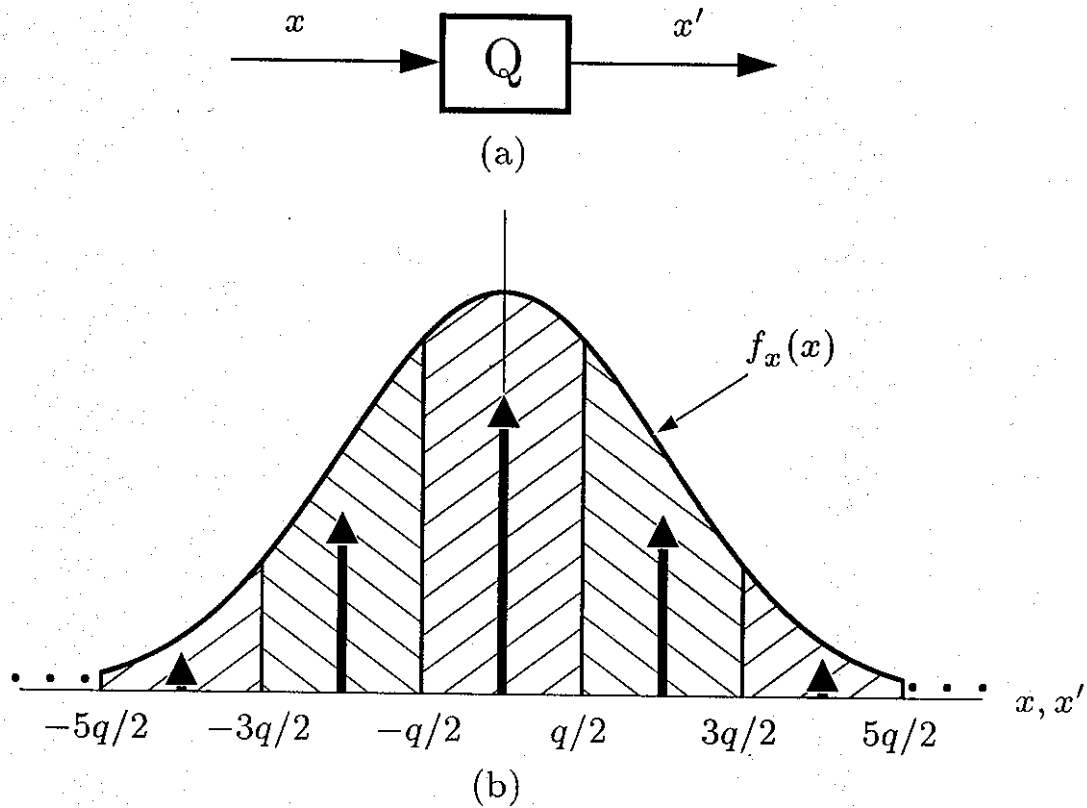


(b)

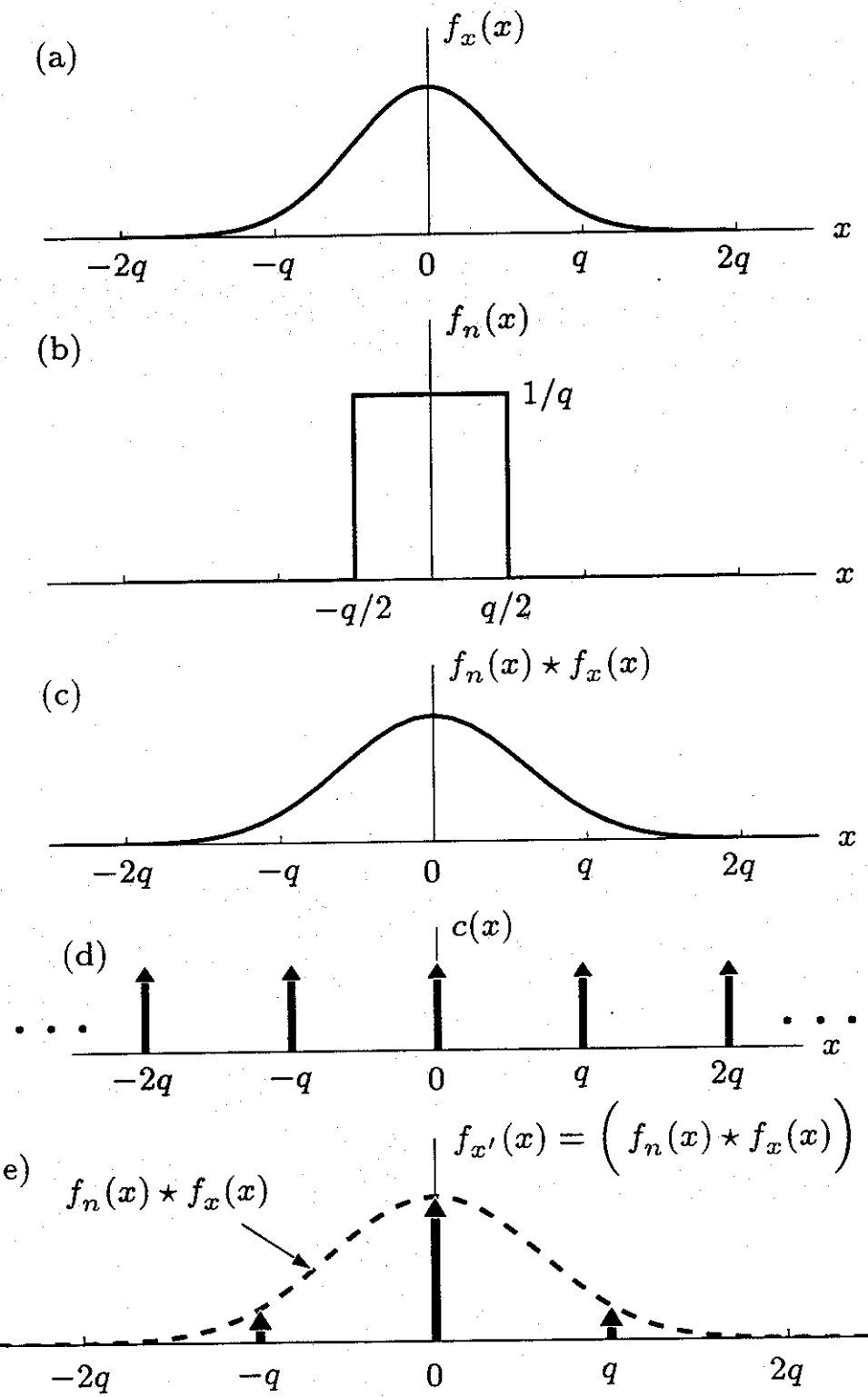
$$\Phi_{x_1,x_2}(u_1,u_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1,x_2)e^{j(u_1x_1+u_2x_2)}\,\mathrm{d}x_1\,\mathrm{d}x_2$$
$$= \mathrm{E}\{e^{ju_1x_1+ju_2x_2}\}. \quad (1)$$

$$\mathrm{E}\{x_1^k x_2^l\} = \frac{1}{j^{k+l}}\left.\frac{\partial^{k+l}\,\Phi_{x_1,x_2}(u_1,u_2)}{\partial u_1^k \partial u_2^l}\right|_{\substack{u_1=0\\u_2=0}}. \quad (2)$$

The Pdf of the Quantizer Output x' "Area Sampling"

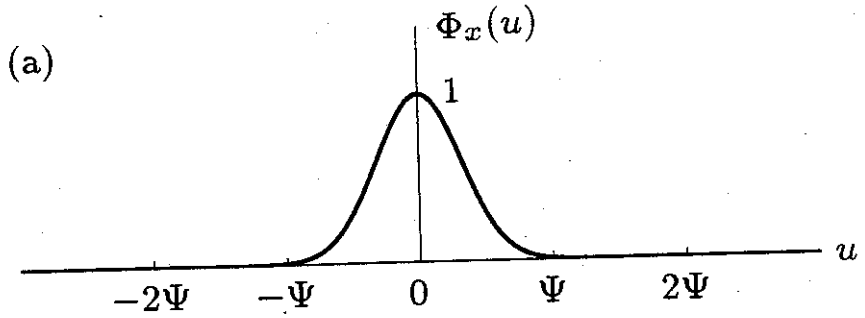


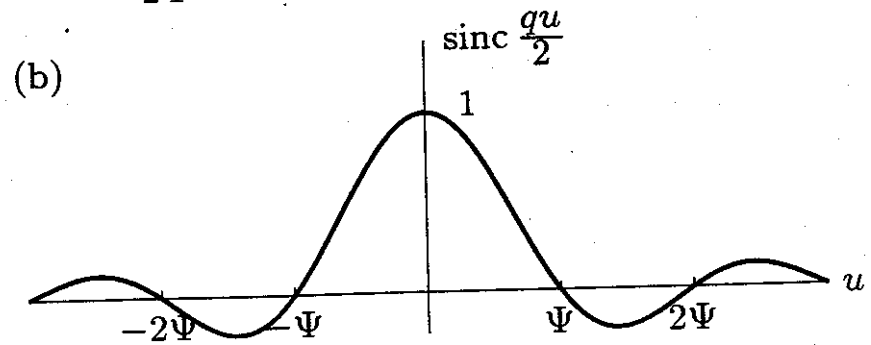
Derivation of Pdf of x' from Area Sampling

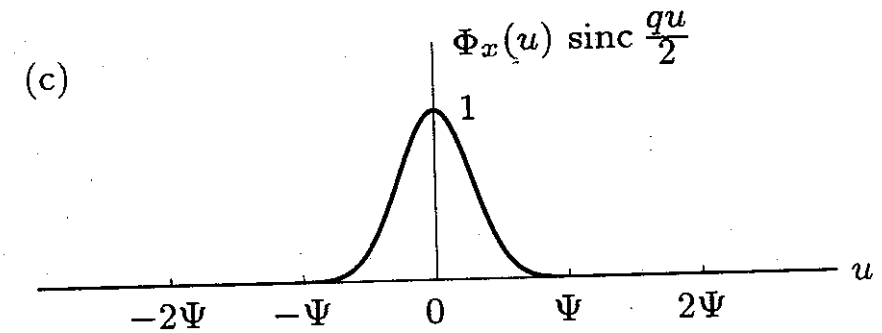


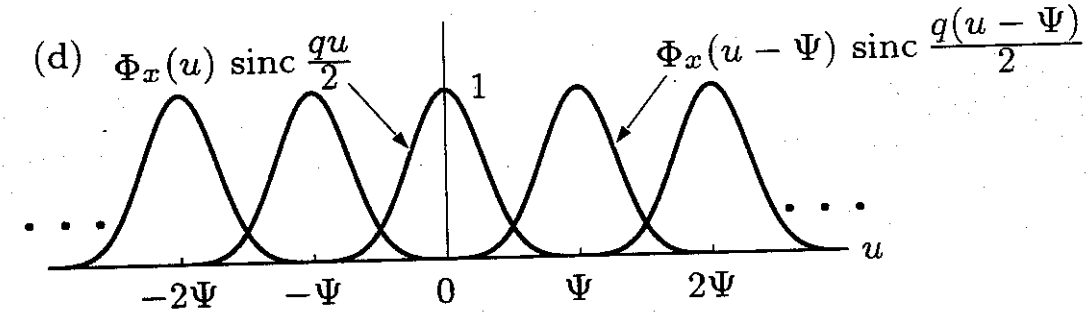
Area Sampling in the Cf Domain

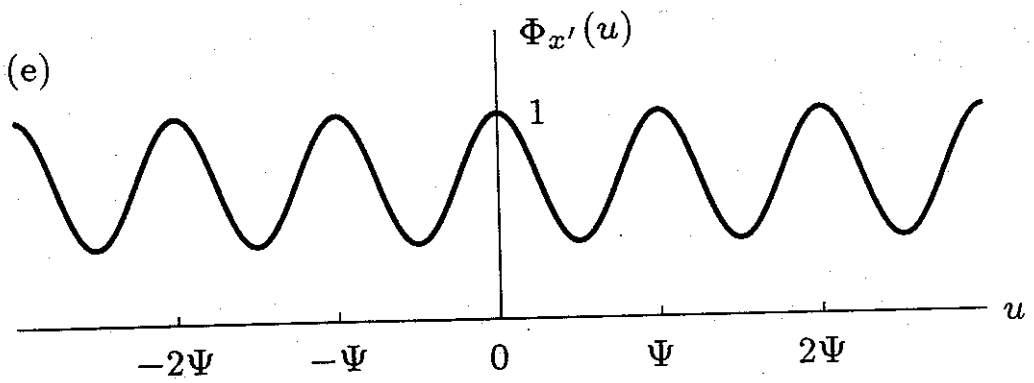
$$\Psi \triangleq \frac{2\pi}{q}$$



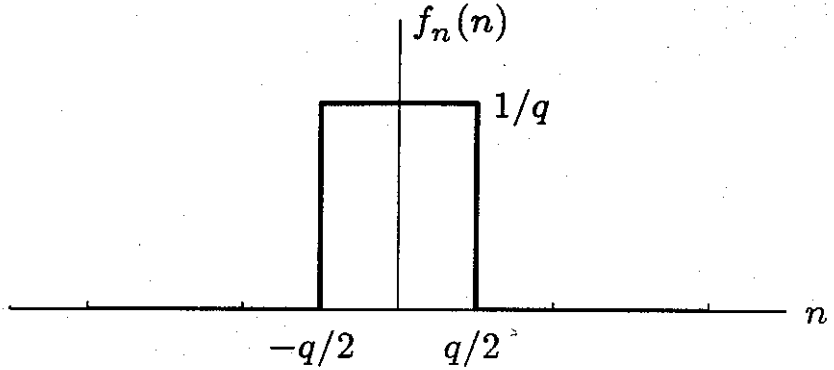


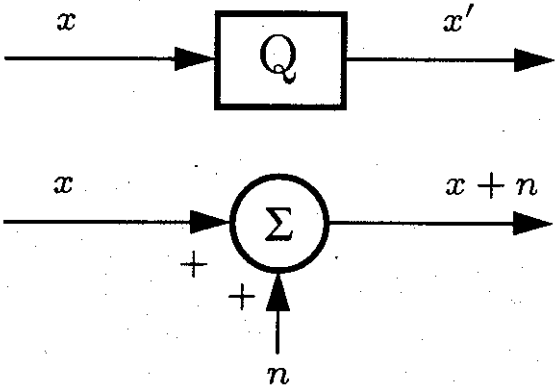




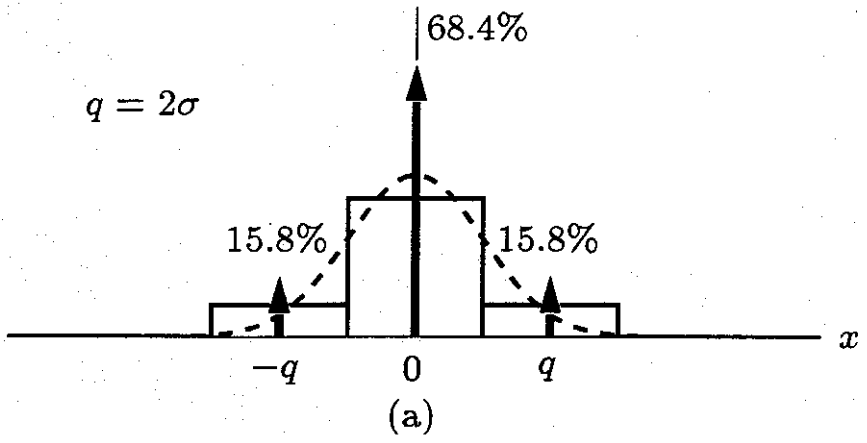


PQN Model

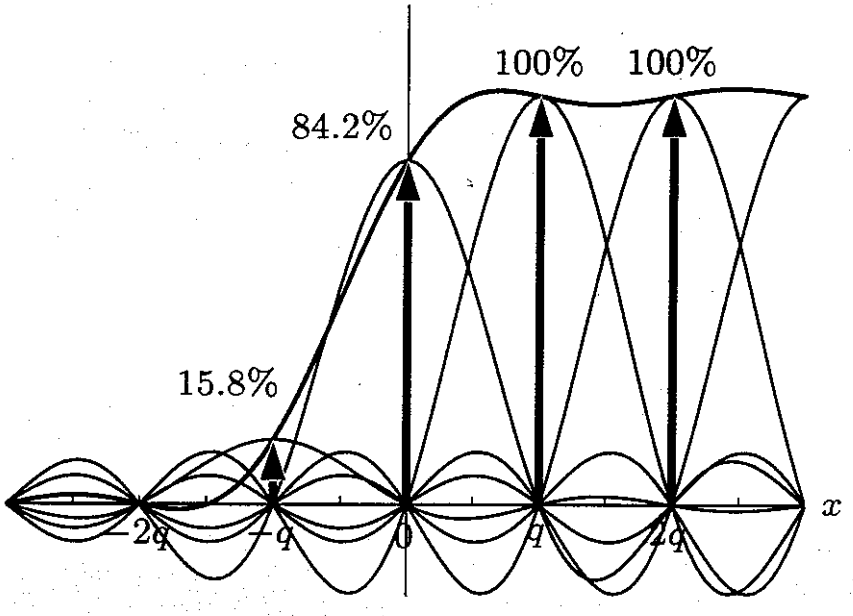




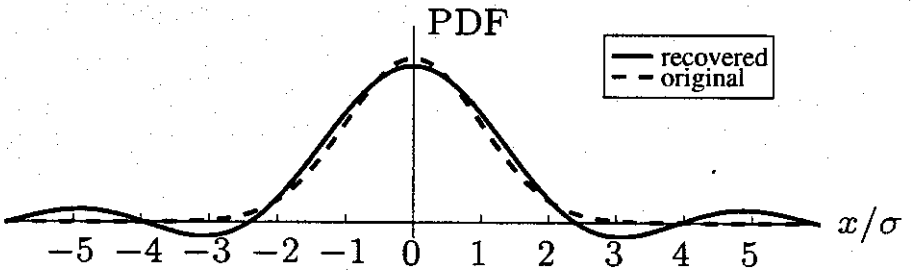
Original Pdf from the Quantized Pdf



(a)

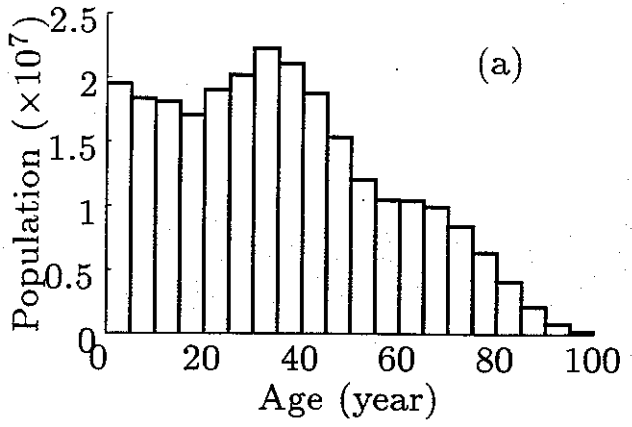


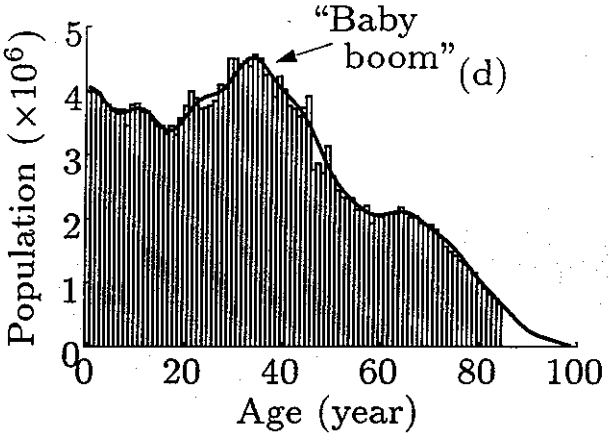
(b)

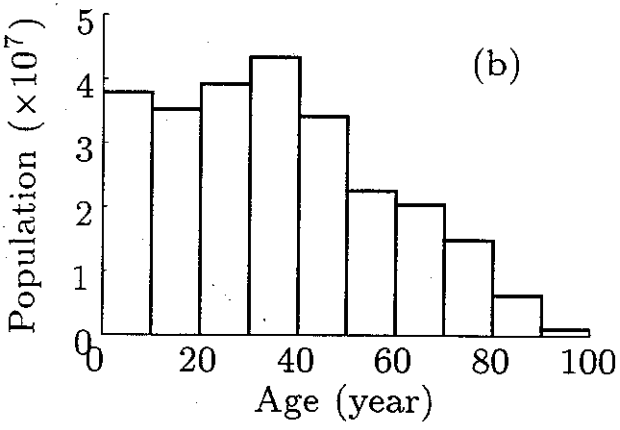


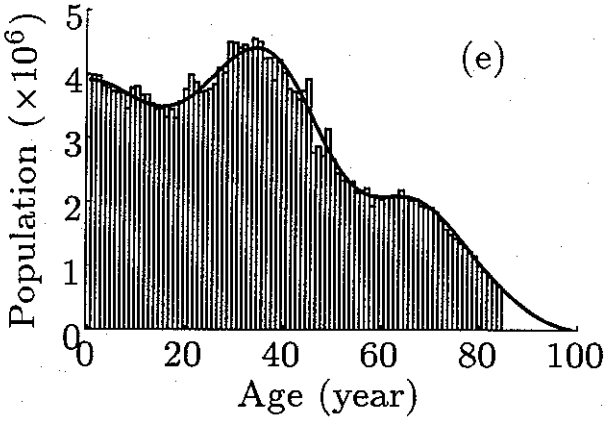
(c)

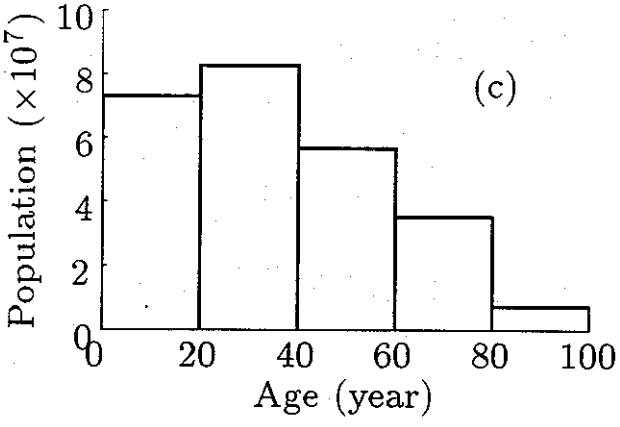
Reconstruction 1990 Census of U. S. Age Distribution

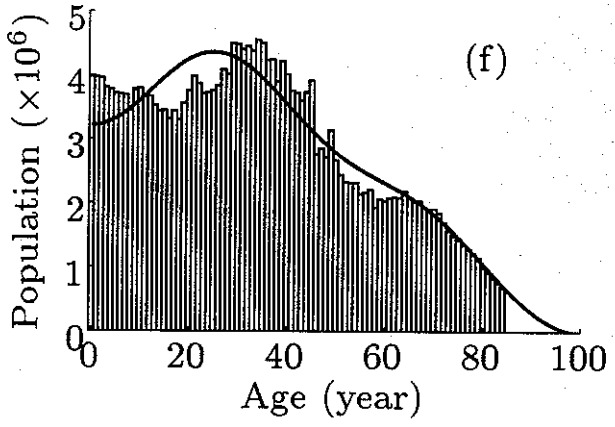




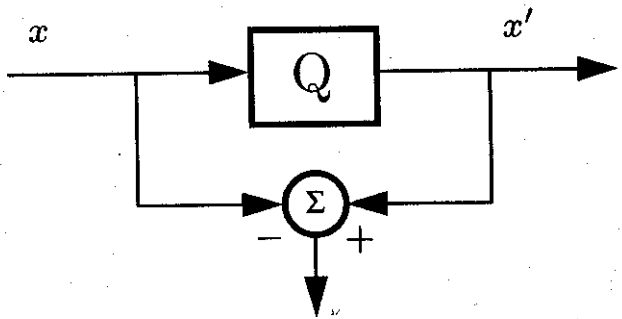




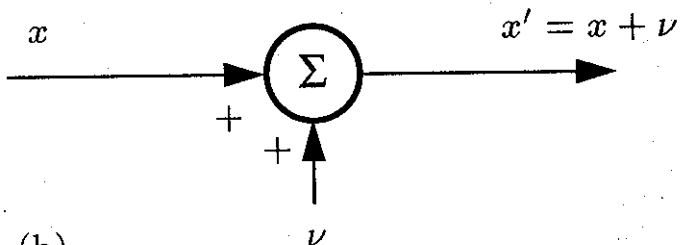




Quantization Noise

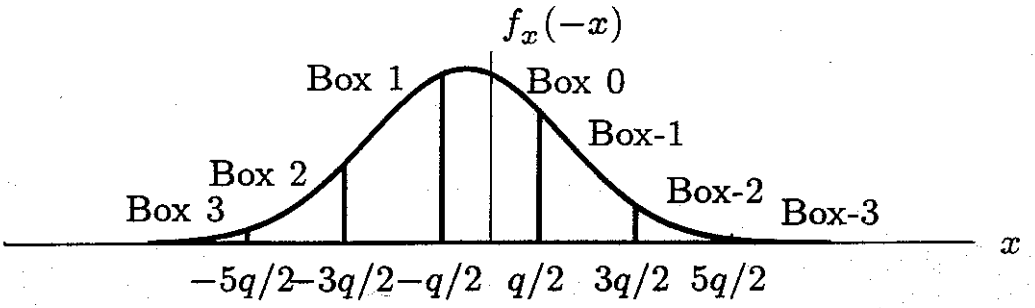


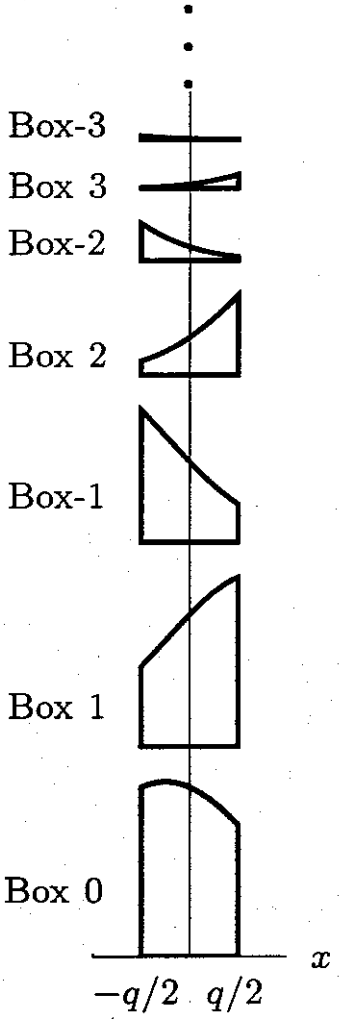
(a)



(b)

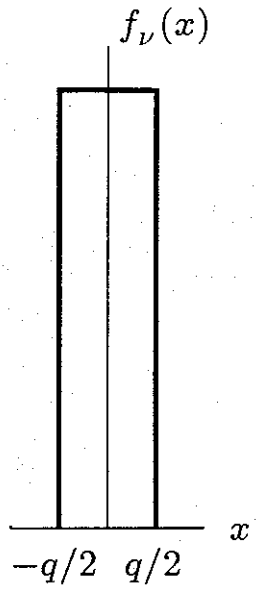
Construction of the Pdf of Quantization Noise



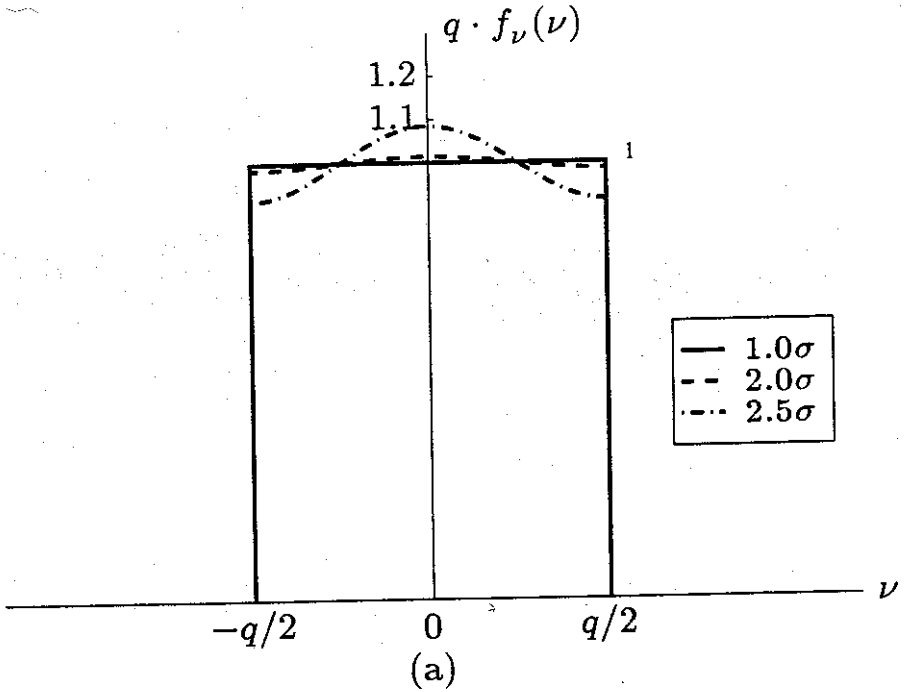


Sum

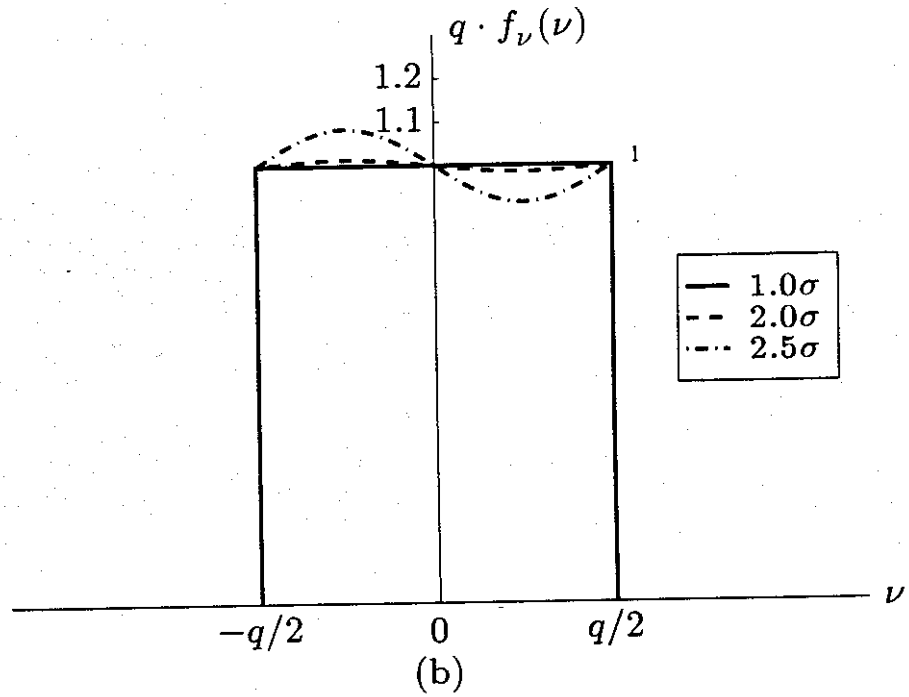
$\Longrightarrow$



The Pdf of Quantization Noise with Gaussian input



(a)



(b)

Moments of the Quantization Noise with Gaussian input

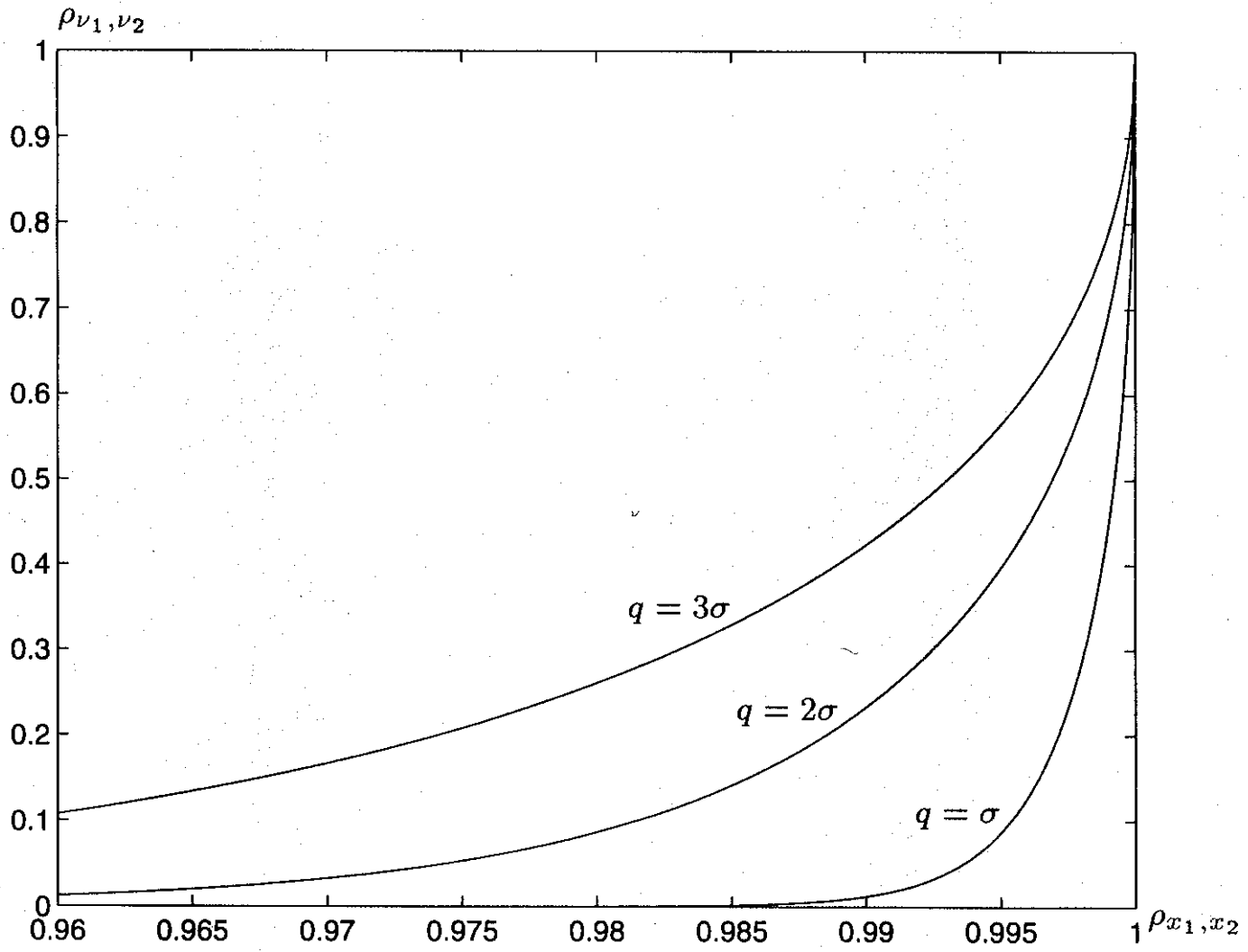
$\mu_x = 0$	$\mathrm{E}\{\nu\}$	$\mathrm{E}\{\nu^2\}$	$\mathrm{E}\{\nu^3\}$	$\mathrm{E}\{\nu^4\}$
$q = 2\sigma$	0	$\left(\frac{1}{12} - 7.3{\cdot}10^{-4}\right) q^2$	0	$\left(\frac{1}{80} - 1.4{\cdot}10^{-4}\right) q^4$
$q = 1.5\sigma$	0	$\left(\frac{1}{12} - 1.6{\cdot}10^{-5}\right) q^2$	0	$\left(\frac{1}{80} - 3.1{\cdot}10^{-6}\right) q^4$
$q = \sigma$	0	$\left(\frac{1}{12} - 2.7{\cdot}10^{-10}\right) q^2$	0	$\left(\frac{1}{80} - 5.3{\cdot}10^{-11}\right) q^4$
$q = 0.5\sigma$	0	$\left(\frac{1}{12} - 5.2{\cdot}10^{-36}\right) q^2$	0	$\left(\frac{1}{80} - 1.0{\cdot}10^{-36}\right) q^4$

$\mu_x = q/4$	$\mathrm{E}\{\nu\}$	$\mathrm{E}\{\nu^2\}$	$\mathrm{E}\{\nu^3\}$	$\mathrm{E}\{\nu^4\}$
$q = 2\sigma$	$-2.3{\cdot}10^{-3} q$	$\left(\frac{1}{12} - 6.8{\cdot}10^{-11}\right) q^2$	$-2.2{\cdot}10^{-4} q^3$	$\left(\frac{1}{80} - 2.9{\cdot}10^{-11}\right) q^4$
$q = 1.5\sigma$	$-4.9{\cdot}10^{-5} q$	$\left(\frac{1}{12} - 1.5{\cdot}10^{-17}\right) q^2$	$-4.8{\cdot}10^{-6} q^3$	$\left(\frac{1}{80} - 6.2{\cdot}10^{-18}\right) q^4$
$q = \sigma$	$-8.5{\cdot}10^{-10} q$	$\left(\frac{1}{12} - 1.7{\cdot}10^{-26}\right) q^2$	$-8.3{\cdot}10^{-11} q^3$	$\left(\frac{1}{80} - 3.3{\cdot}10^{-27}\right) q^4$
$q = 0.5\sigma$	$-1.6{\cdot}10^{-35} q$	$\left(\frac{1}{12} - 3.2{\cdot}10^{-52}\right) q^2$	$-1.6{\cdot}10^{-36} q^3$	$\left(\frac{1}{80} - 6.2{\cdot}10^{-53}\right) q^4$

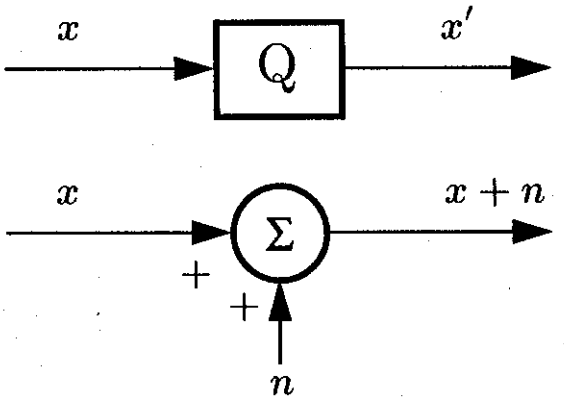
Correlation Coefficients between Gaussian input and Noise

	$\mu_x = 0, \quad \rho_{x,\nu} = \dfrac{\mathrm{cov}\{x,\nu\}}{\sqrt{\mathrm{var}\{x\}\,\mathrm{var}\{\nu\}}}$	$\mu_x = q/4, \quad \rho_{x,\nu} = \dfrac{\mathrm{cov}\{x,\nu\}}{\sqrt{\mathrm{var}\{x\}\,\mathrm{var}\{\nu\}}}$
$q = 2\sigma$	-2.50×10^{-2}	-1.77×10^{-2}
$q = 1.5\sigma$	-7.15×10^{-4}	-3.58×10^{-4}
$q = \sigma$	-1.85×10^{-8}	-1.07×10^{-24}
$q = 0.5\sigma$	-7.10×10^{-34}	7.10×10^{-34}

Relationship between ρ_{ν_1,ν_2} and ρ_{x_1,x_2}



Comparison of Quantization and PQN



$$\Phi_{x,\nu,x'}(u_x, u_\nu, u_{x'}) = \sum_{l=-\infty}^{\infty} \Phi_{x,n,x+n}(u_x, u_\nu, u_{x'} + l\Psi)\,. \tag{3}$$

This is the fundamental relation between the CF's of quantization and PQN. What Eq. (3) tells us is that the three-dimensional CF for quantization is periodic along the $u_{x'}$-axis, and aperiodic along the u_x and u_ν-axes. If we could draw it, we would see that it is an infinite sum of replicas of the three-dimensional CF of PQN displaced by integer multiples of Ψ along the $u_{x'}$-axis. Recall that Ψ is the quantization "radian frequency," equal to $\Psi = 2\pi/q$. Periodicity of the CF results from the fact that x' can only exist at uniformly spaced discrete levels.

Joint Cf for Quantization

$$\Phi_{x_1,\ldots,x_N,\nu_1,\ldots,\nu_N,x'_1,\ldots,x'_N}(u_{x_1},\ldots,u_{x_N},u_{\nu_1},\ldots,u_{\nu_N},u_{x'_1},\ldots,u_{x'_N})$$

$$= \sum_{l_1=-\infty}^{\infty} \cdots \sum_{l_N=-\infty}^{\infty} \Phi_{x_1,\ldots,x_N,n_1,\ldots,n_N,x_1+n_1,\ldots,x_N+n_N}(u_{x_1},\ldots,u_{x_N},$$

$$u_{\nu_1},\ldots,u_{\nu_N},u_{x'_1}+l_1\Psi_1,\ldots,u_{x'_N}+l_N\Psi_N)\,. \qquad (4)$$

Uniform Quantization Summary

If QT II is satisfied, the PQN model applies:

quantization of one variable

- ν is uniformly-distributed
- $\mathrm{E}\{\nu\} = 0$
- $\mathrm{E}\{\nu^2\} = q^2/12$
- $\mathrm{cov}\{\nu\} = 0$

quantization of two variables

- all of the above applies to each of the variables
- $\mathrm{E}\{\nu_1 \nu_2\} = 0$

quantization of three or more variables

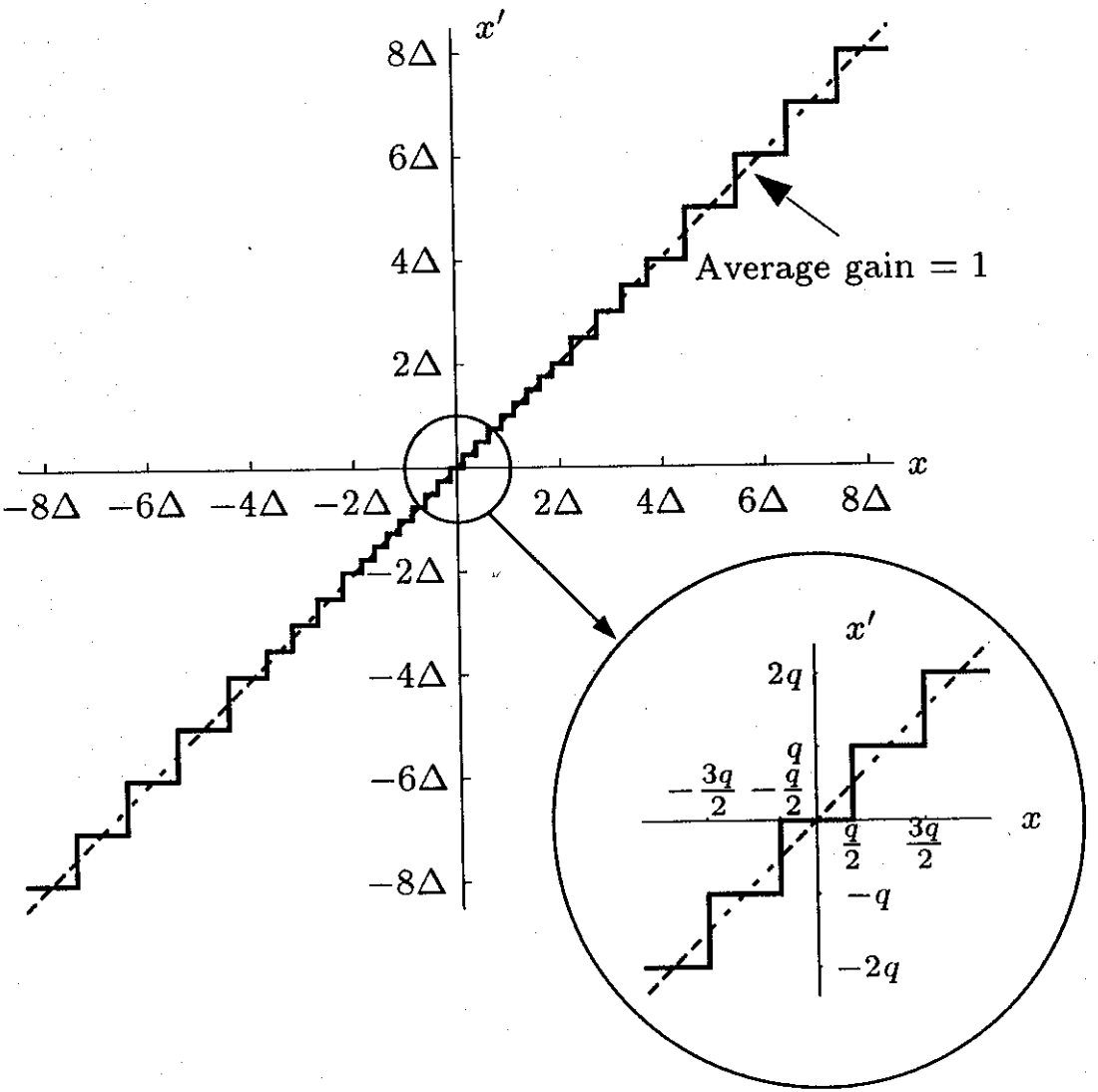
- $\mathrm{E}\{\nu_1 \ldots \nu_N\} = 0$

If PQN model applies, the quantizer may be replaced for purpose of analysis by a source of additive independent white noise.

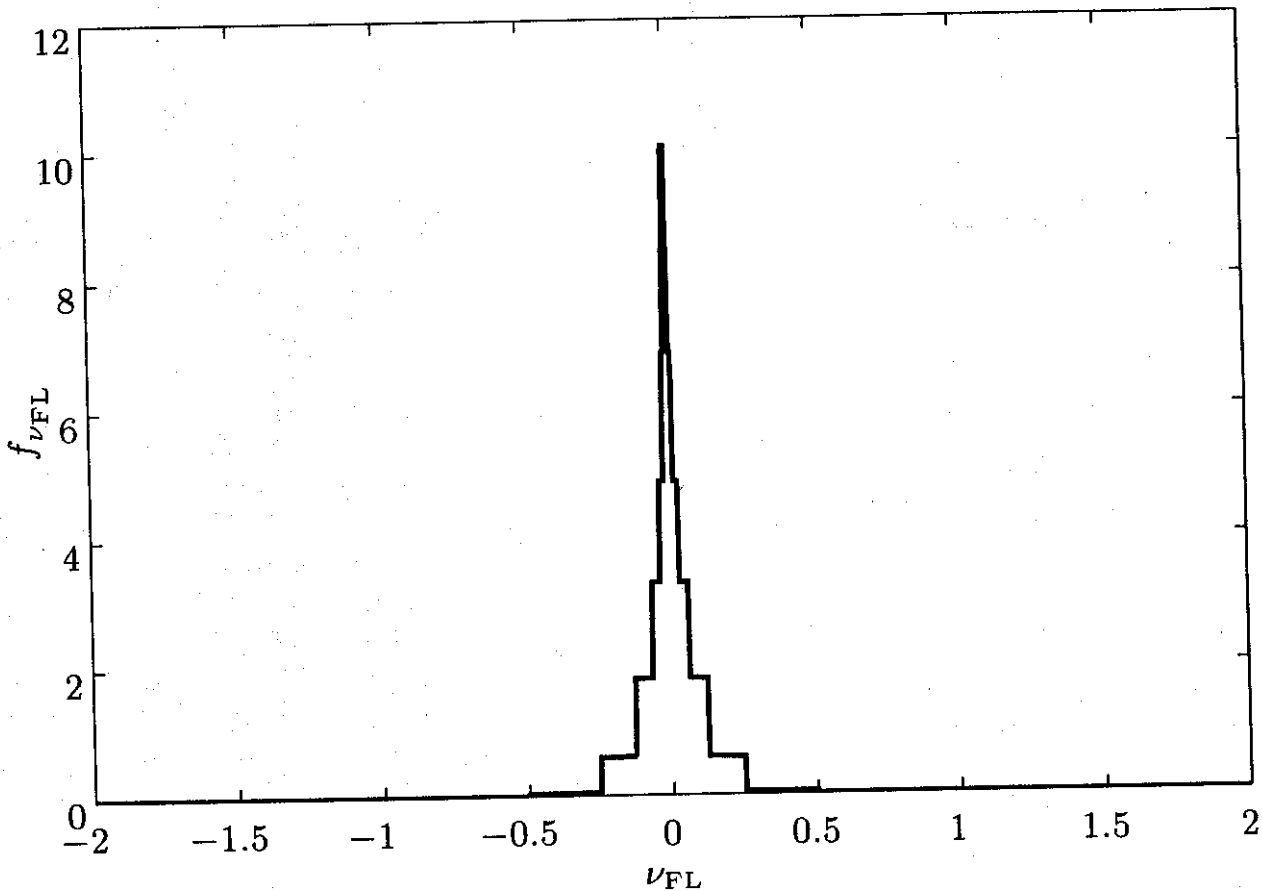
Counting with Binary Floating-Point numbers with 4-bit Mantissa

	Mantissa	
0	0000	
1	0001	
2	0010	
3	0011	
4	0100	
5	0101	
6	0110	
7	0111	
8	1000	$\times 2^{E}$
9	1001	
10	1010	
11	1011	
12	1100	
13	1101	
14	1110	
15	1111	

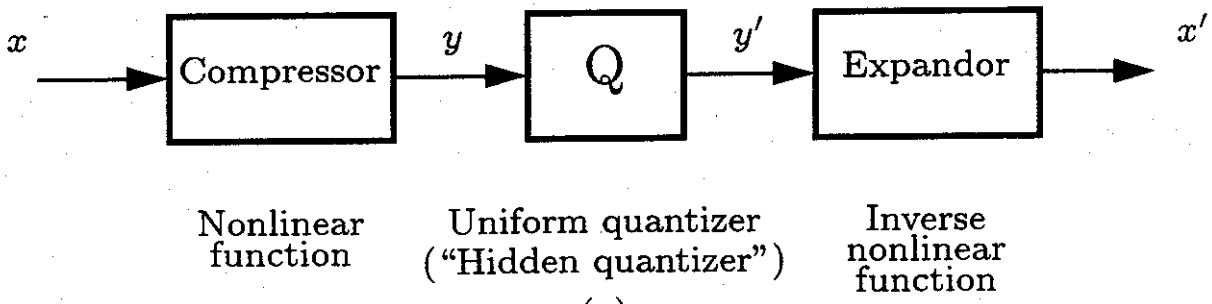
A Floating-Point Quantizer with 3-bit Mantissa



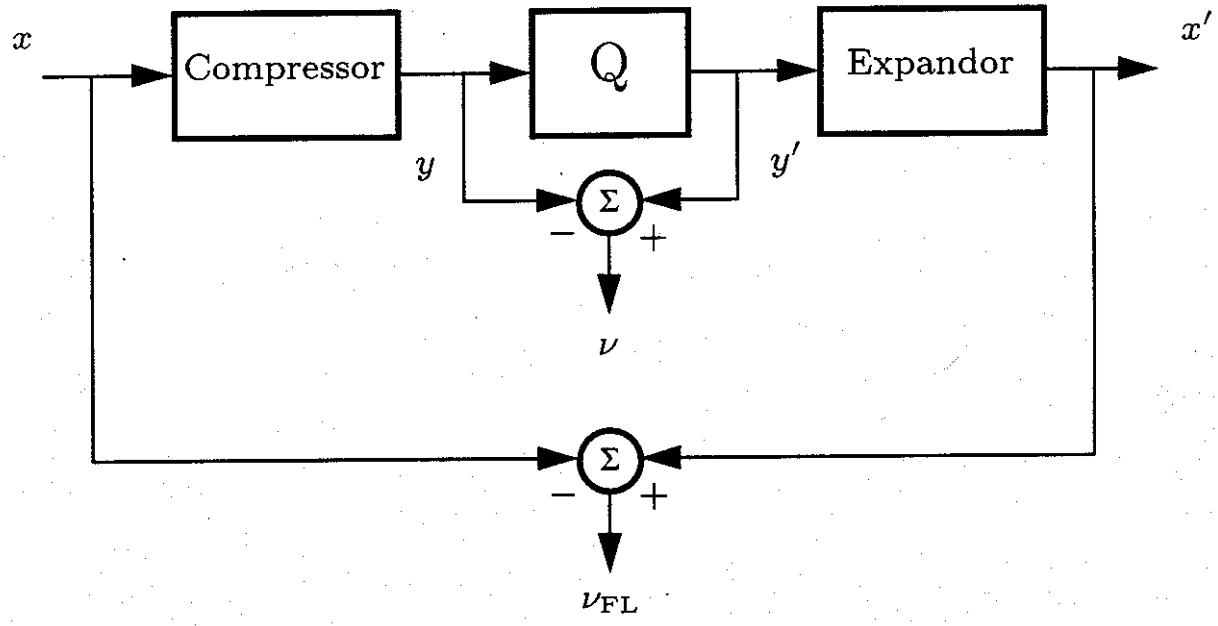
Pdf of Floating-Point Quantization Noise



A Model of a Floating-Point Quantizer

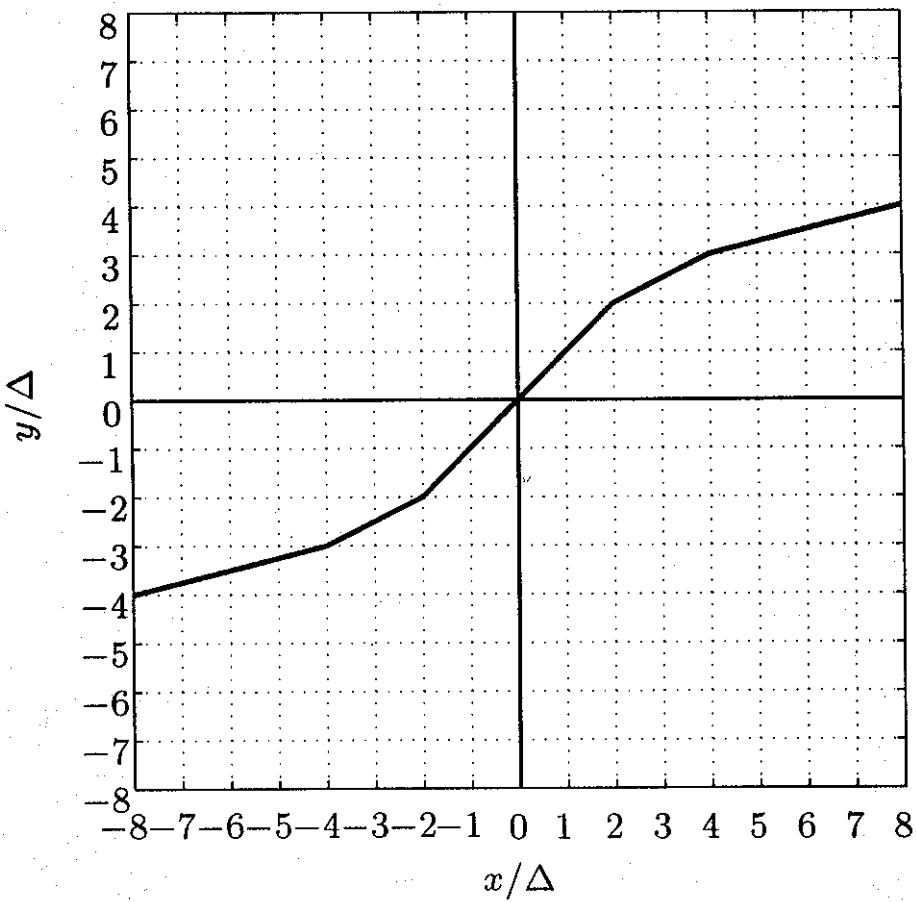


(a)

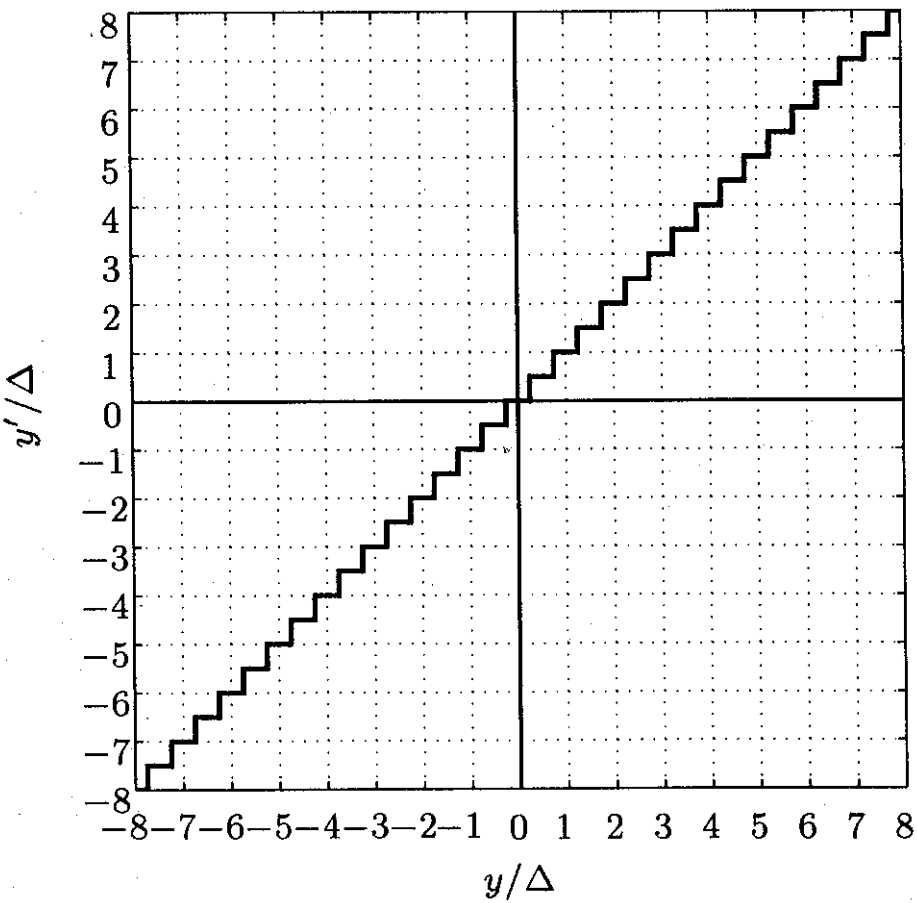


(b)

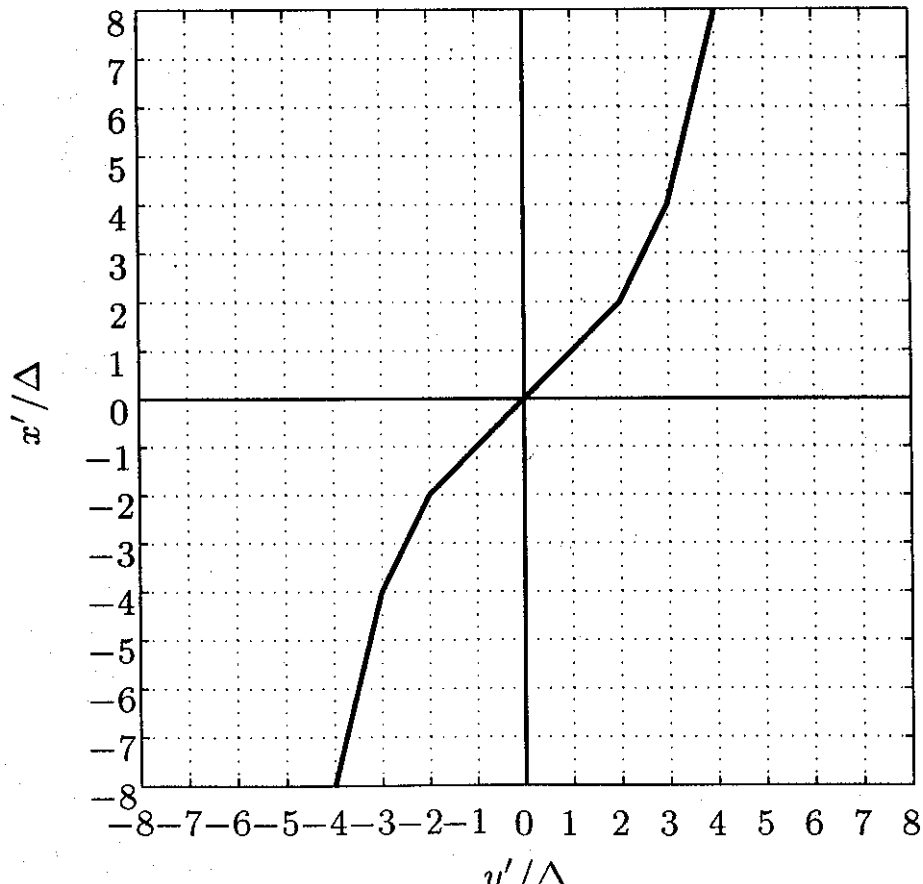
The Compressor's Input-Output Characteristic



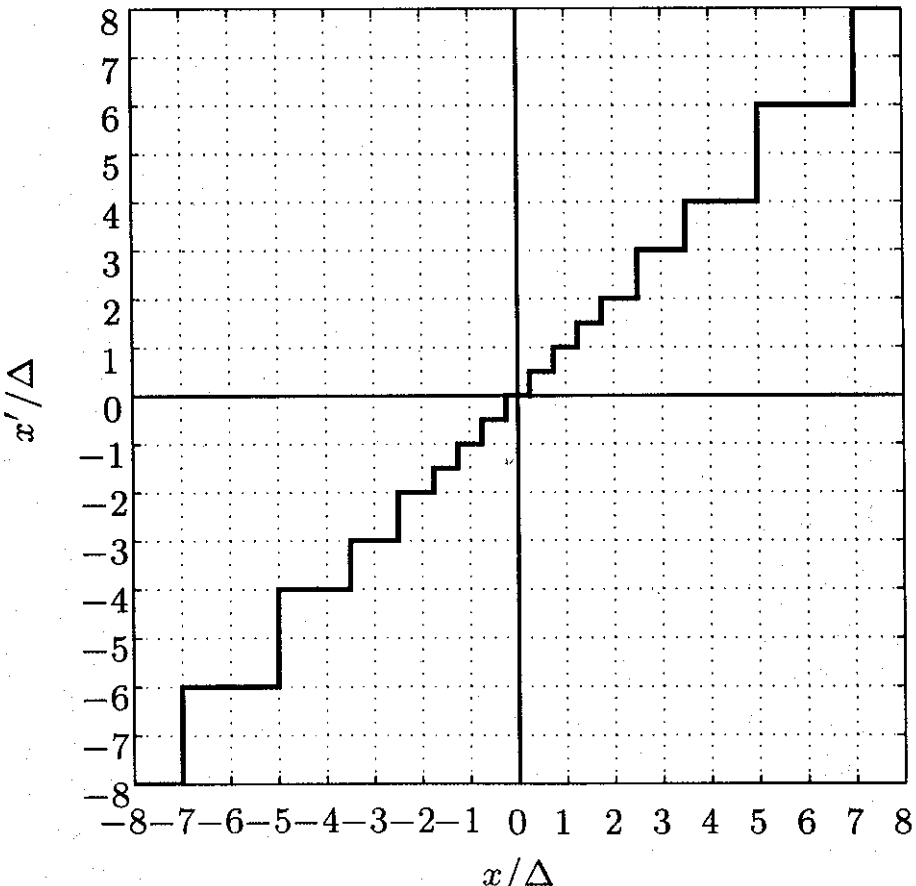
The Uniform "Hidden Quantizer" with 2-bit Mantissa



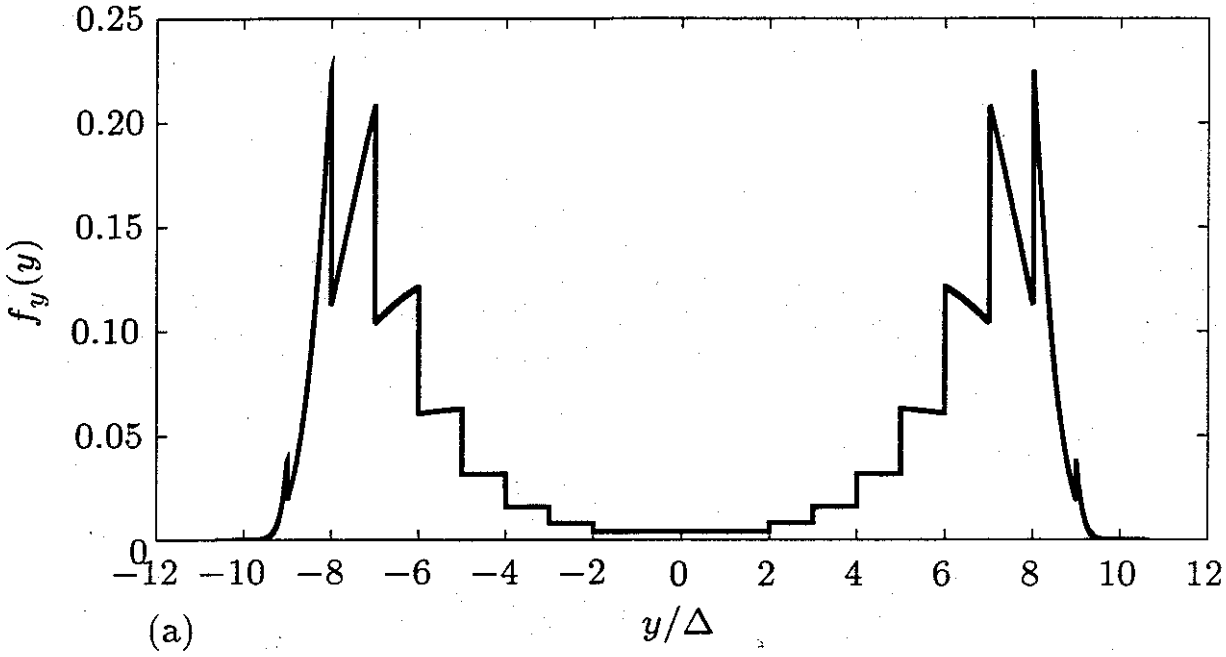
The Expandor's Input-Output Characteristic



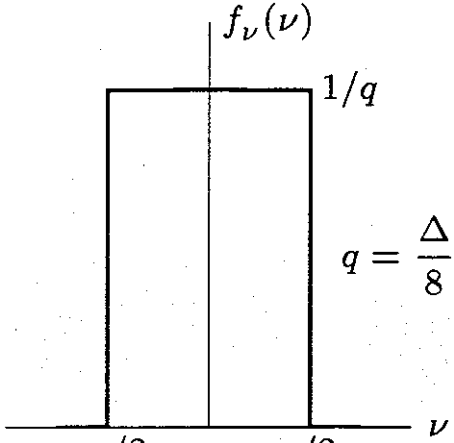
A Floating-Point Quantizer with 2-bit Mantissa



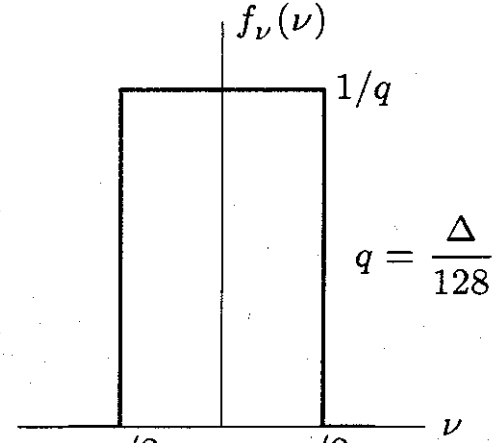
Compressor Output and Hidden Quantization Noise



(a)



(b)



(c)

The input is zero-mean Gaussian with $\sigma = 100\Delta$.

Floating-Point Quantization Summary

When PQN model applies to "hidden quantizer":

quantization of one variable

- ν_{FL} is "skyscraper-distributed"
- $\mathrm{E}\{\nu_{\mathrm{FL}}\} = 0$
- $(\frac{1}{12})2^{-2p}\mathrm{E}\{x^2\} \leq \mathrm{E}\{\nu_{\mathrm{FL}}^2\} \leq (\frac{1}{3})2^{-2p}\mathrm{E}\{x^2\}$
- $\mathrm{E}\{\nu_{\mathrm{FL}}^2\} \approx 0.180 \times 2^{-2p}\mathrm{E}\{x^2\}$
- $SNR \triangleq \frac{\mathrm{E}\{x^2\}}{\mathrm{E}\{\nu_{\mathrm{FL}}^2\}}$
- $3 \times 2^{2p} \leq SNR \leq 12 \times 2^{2p}$
- $SNR \approx 5.55 \times 2^{2p}$
- with 16-bit mantissa,
  $SNR \approx 5.55 \times 2^{2p} = 2.38 \times 10^{10}$ (104 dB)
- $\mathrm{cov}\{x\nu_{\mathrm{FL}}\} = 0$

quantization of two variables

- all of the above applies to each of the variables
- $\mathrm{E}\{\nu_{\mathrm{FL}_1}\nu_{\mathrm{FL}_2}\} = 0$

quantization of three or more variables

- $\mathrm{E}\{\nu_{\mathrm{FL}_1} \ldots \nu_{\mathrm{FL}_N}\} = 0$

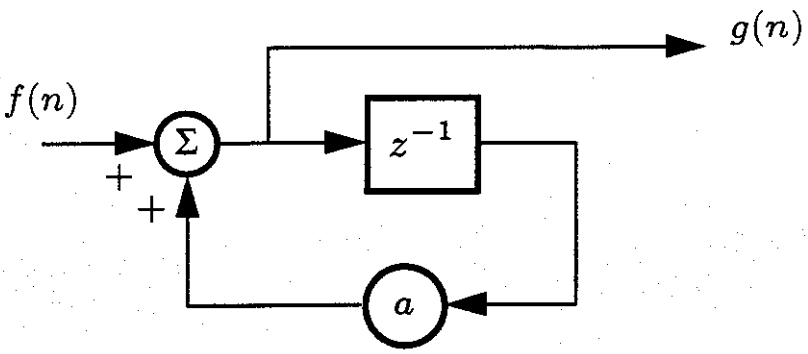
When PQN model applies to hidden quantizer, the floating-point quantizer may be replaced for purpose of analysis by a source of additive independent white noise.

PQN Model for "Hidden Quantizer" Works!

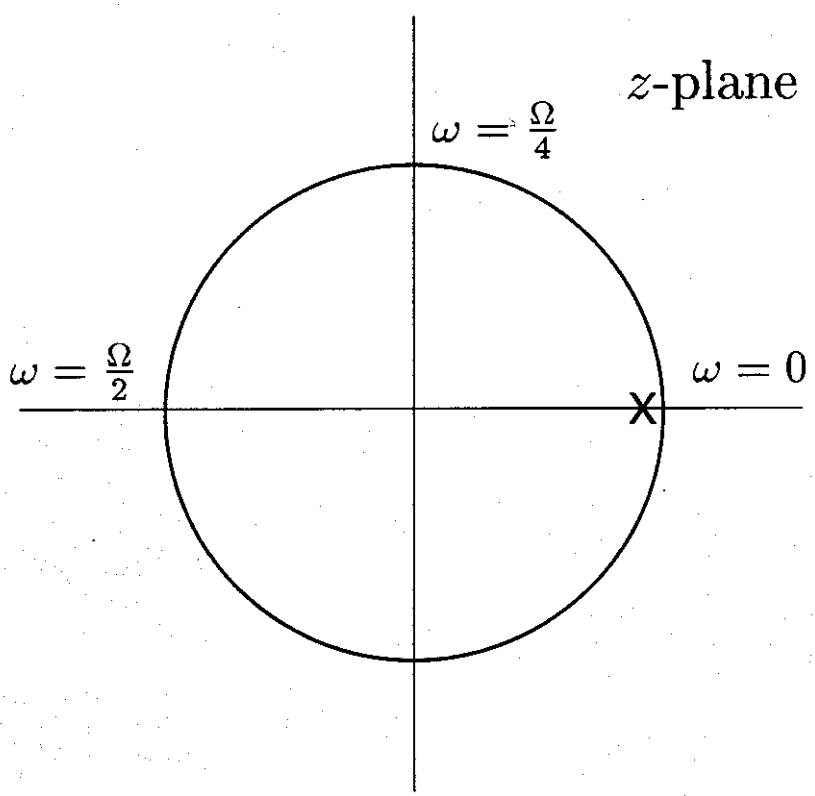
	Mean of noise	Normalized mean square of quantization noise $\frac{\mathrm{E}\{\nu_{\mathrm{FL}}^2\}}{\mathrm{E}\{x^2\}} \times 2^{2p}$	Correlation coefficient of ν_{FL} and x	Correlation coefficient of ν_{FL_1} and ν_{FL_2}	
				$\rho_{x_1,x_2} = 0.99$	$\rho_{x_1,x_2} = 0.999999$
gaussian	0	0.181	3×10^{-3}	$< 10^{-6}$	$< 10^{-3}$
triangular-distributed	0	0.189	6×10^{-4}		
rectangular-distributed	0	0.199	6×10^{-4}		
sinusoidal	0	0.166	1×10^{-2}	10^{-2}	10^{-1}

Zero-mean input and 16-bit mantissa for floating-point quantizer.

Roundoff in Digital Filters



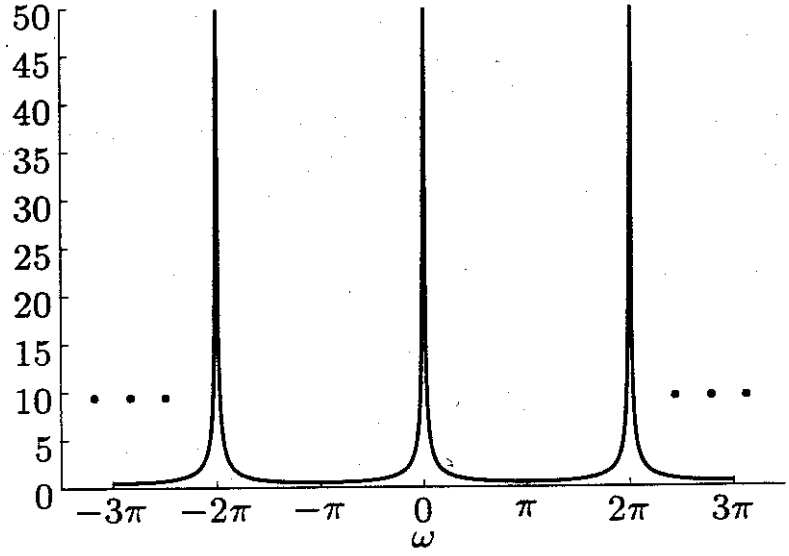
$a = 0.98$



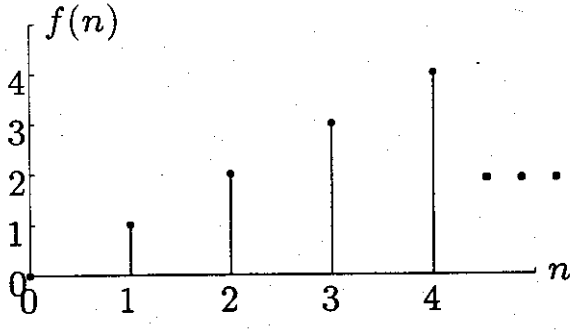
$$H(z) = \frac{G(z)}{F(z)} = \frac{1}{1 - 0.98z^{-1}}$$

Sinusoidal response

$$\omega = 0, \quad |gain| = 50$$
$$\omega = \tfrac{\Omega}{2}, \quad |gain| \approx 0.5$$



Let input be $f(n) = n$

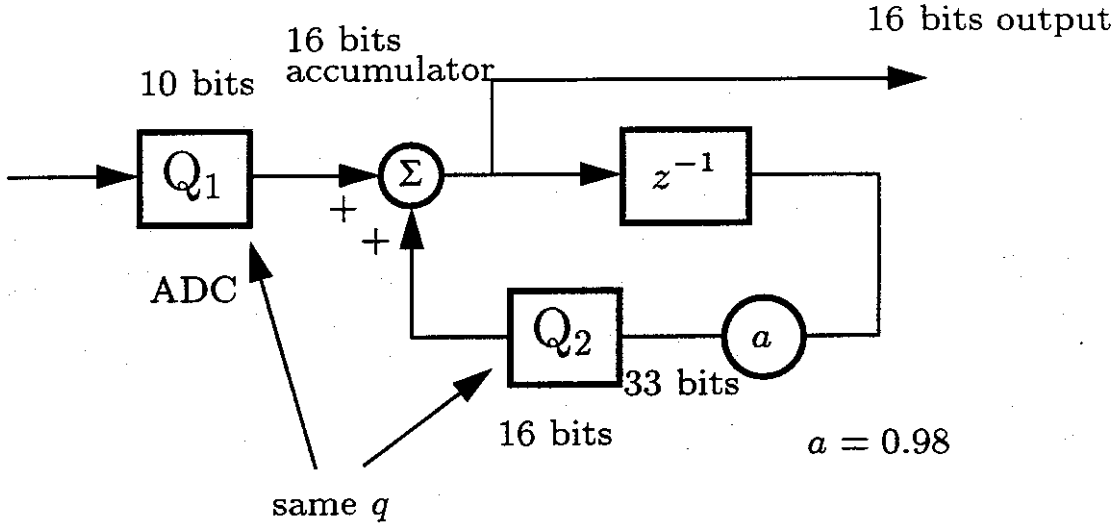


$$g(n) = f(n) + ag(n-1)$$

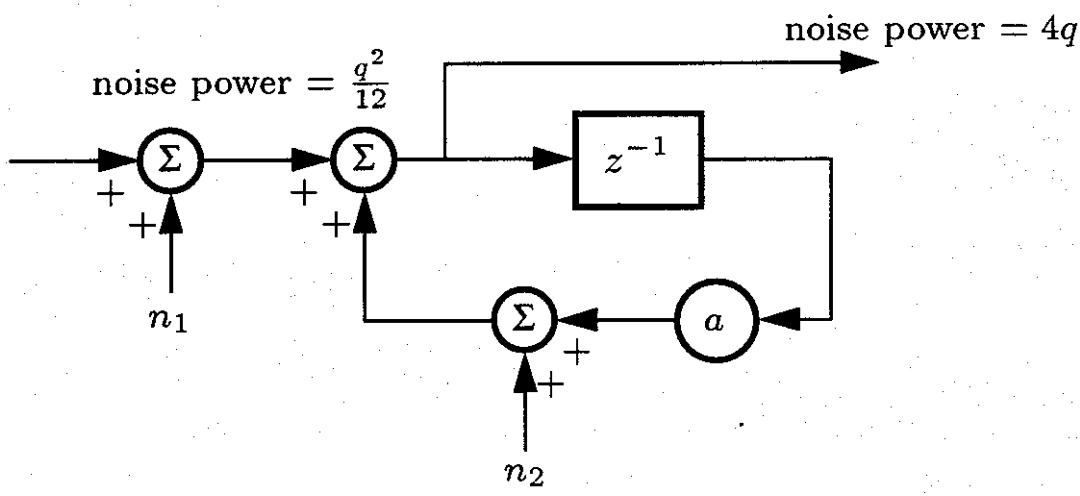
$$g(0) = 0 + 0.98(0) = 0$$
$$g(1) = 1 + 0.98(0) = 1$$
$$g(2) = 2 + 0.98(1) = 2.98$$
$$g(3) = 3 + 0.98(2.98)$$
$$g(4) = 4 + 0.98(3 + 0.98(2.98))$$

$\cdots$ word length grows!

Implementation:



Analysis:



Impulse response from n_1 or n_2 to output is $1, 0.98, 0.98^2, 0.98^3, \ldots$

total output quantization noise power

$$= 2\frac{q^2}{12}\left[1 + 0.98^2 + 0.98^4 + \cdots\right]$$
$$= \frac{q^2}{6}\frac{1}{1 - 0.98^2} \approx 4q^2$$

noise standard deviation at output $\approx 2q$

Let input be sine wave of amplitude A, let $\pm A$ be full scale —

signal power $= A^2/2$
noise power $= q^2/12$
input SNR $= \frac{512^2 q^2/2}{q^2/12} = 6 \times 512^2$
at filter output, noise power $= 4q^2$

input low frequency sine, amplitude A

$$\text{output signal mag.} = 50A$$

$$\text{output signal power} = 50^2 \frac{A^2}{2}$$

$$\text{output SNR} = \frac{50^2 A^2/2}{4q^2} = \frac{50^2 512^2 q^2/2}{4q^2}$$
$$= \frac{50^2}{8} 512^2 = 312.5 \times 512^2$$

input high frequency sine, amplitude A

$$\text{output signal mag.} = A/2$$

$$\text{output signal power} = \frac{A^2/4}{2} = A^2/8$$

$$\text{output SNR} = \frac{A^2/8}{4q^2} = \frac{1}{32} 512^2$$

worst case output SNR, poorer than input SNR.