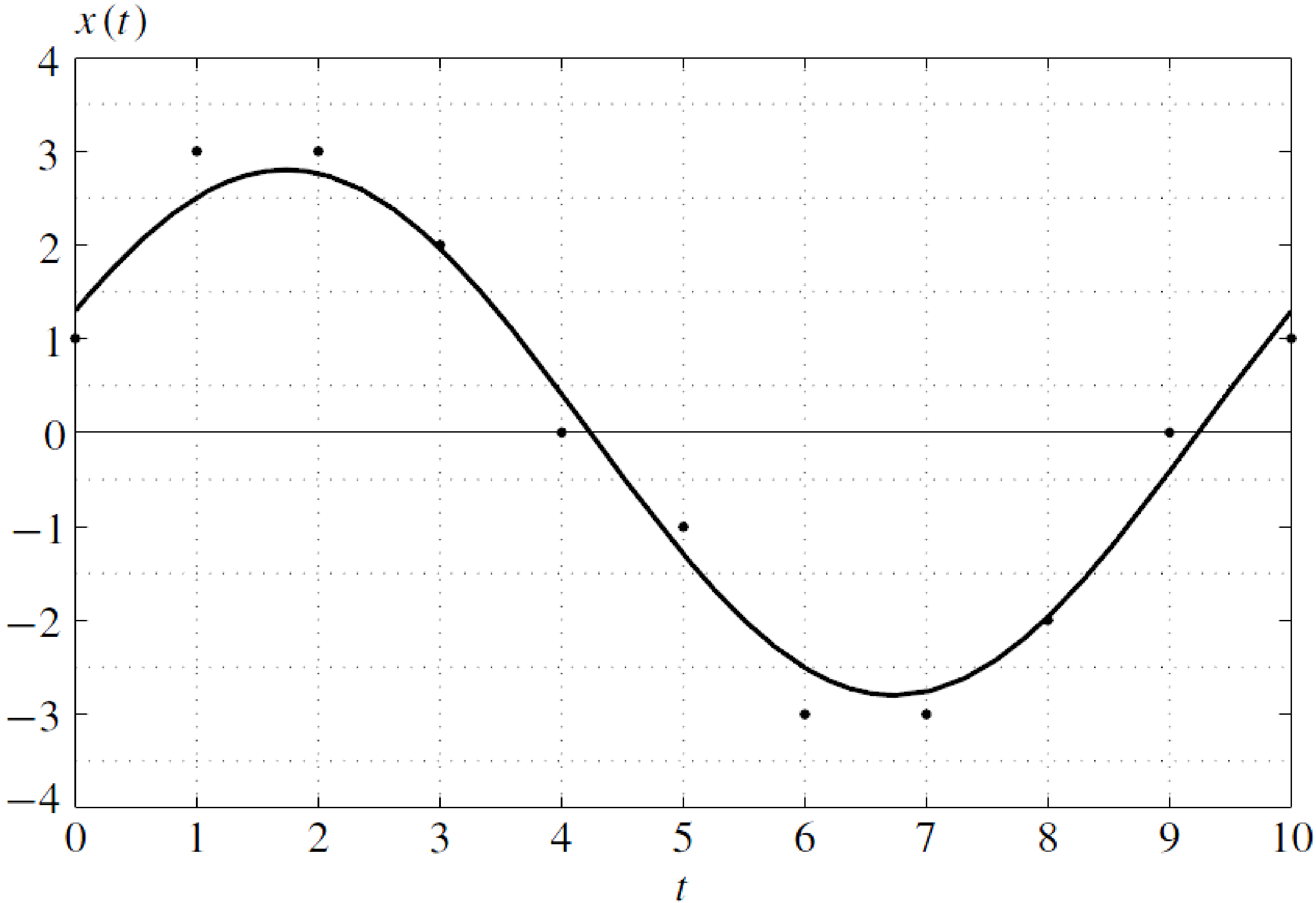
Quantization Noise

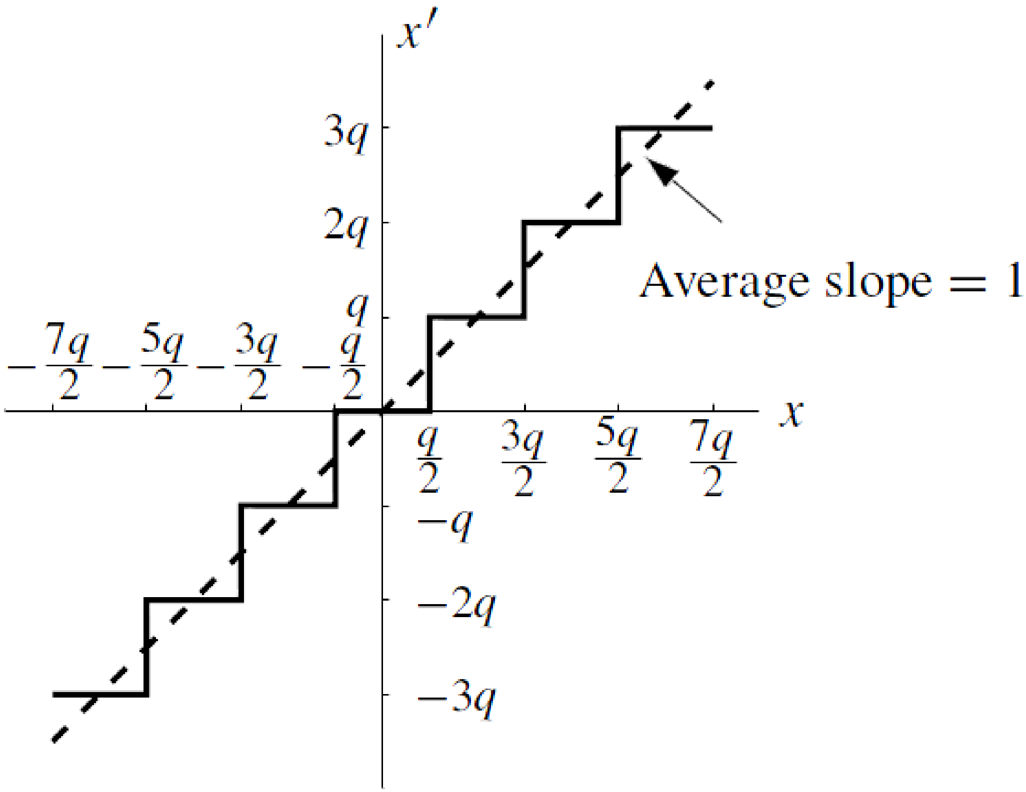
Bernard Widrow

Information Systems Laboratory
Department of Electrical Engineering
Stanford University

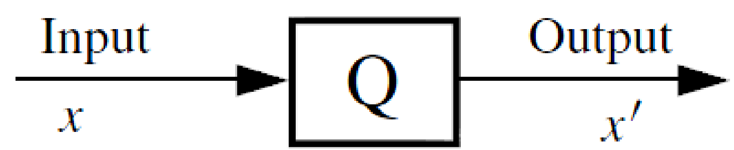
Sampling and Quantization



A Basic Quantizer

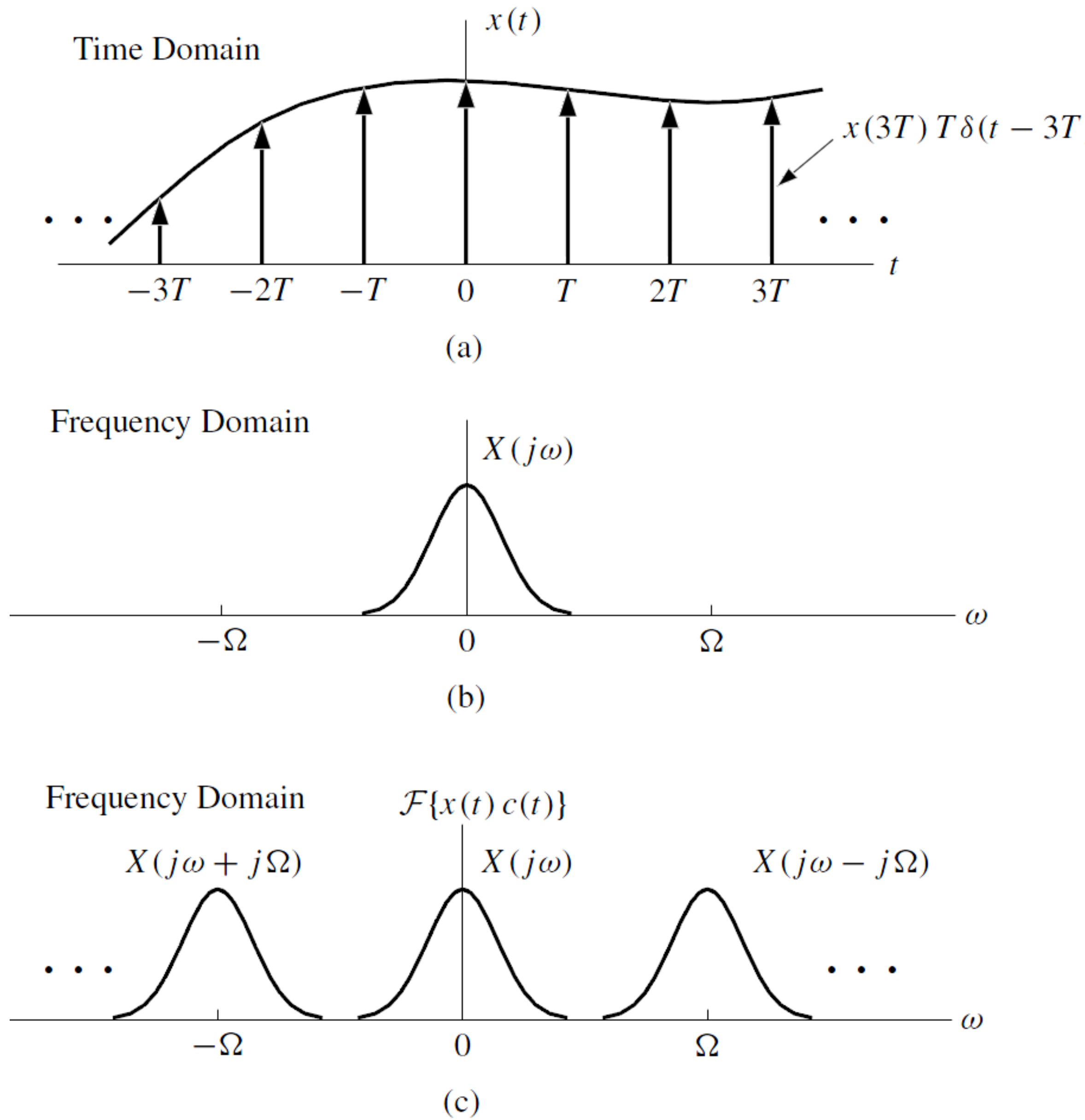


(a)

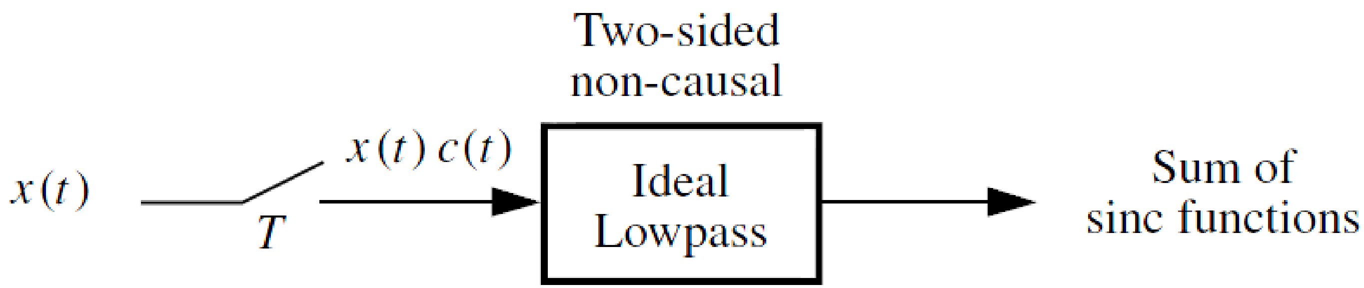


(b)

Fourier Transform of a Sampled Function

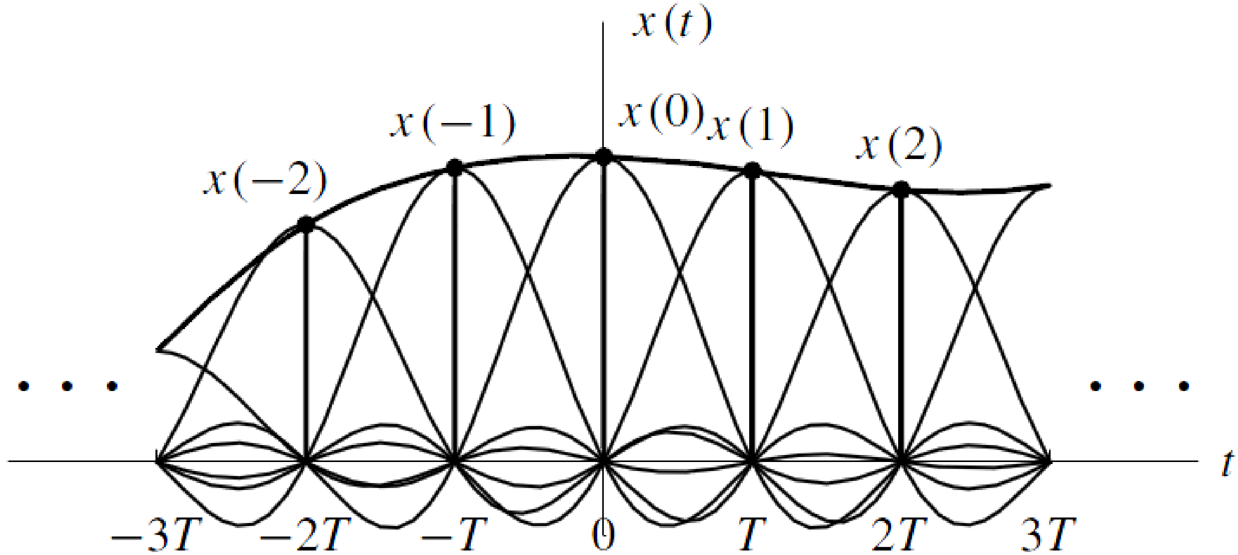


Recovery of Original Signal



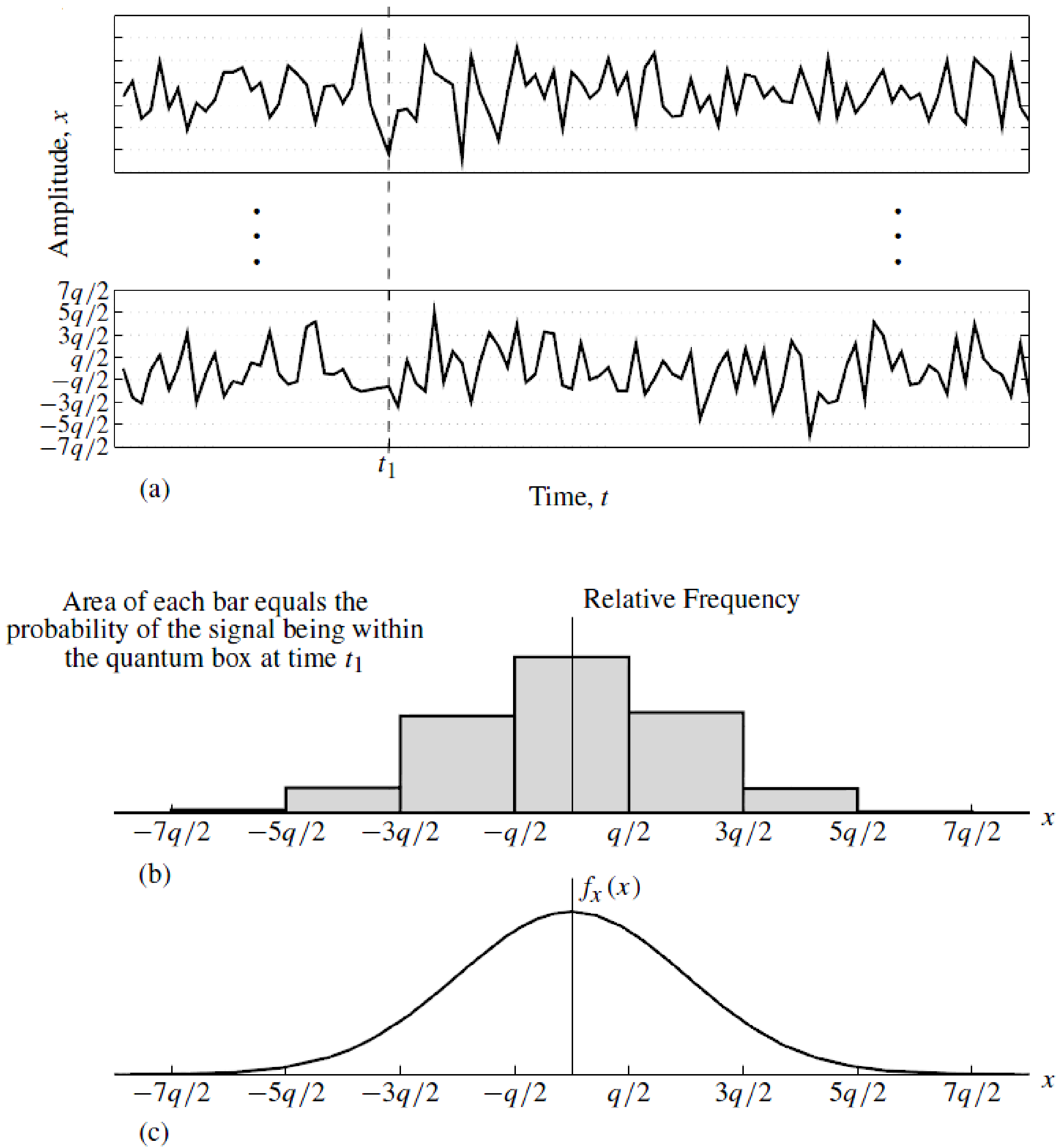
$$\frac{1}{T}\,\mathrm{sinc}\left(\frac{\pi t}{T}\right) = \frac{1}{T}\,\frac{\sin\left(\frac{\pi t}{T}\right)}{\frac{\pi t}{T}}$$

(a)

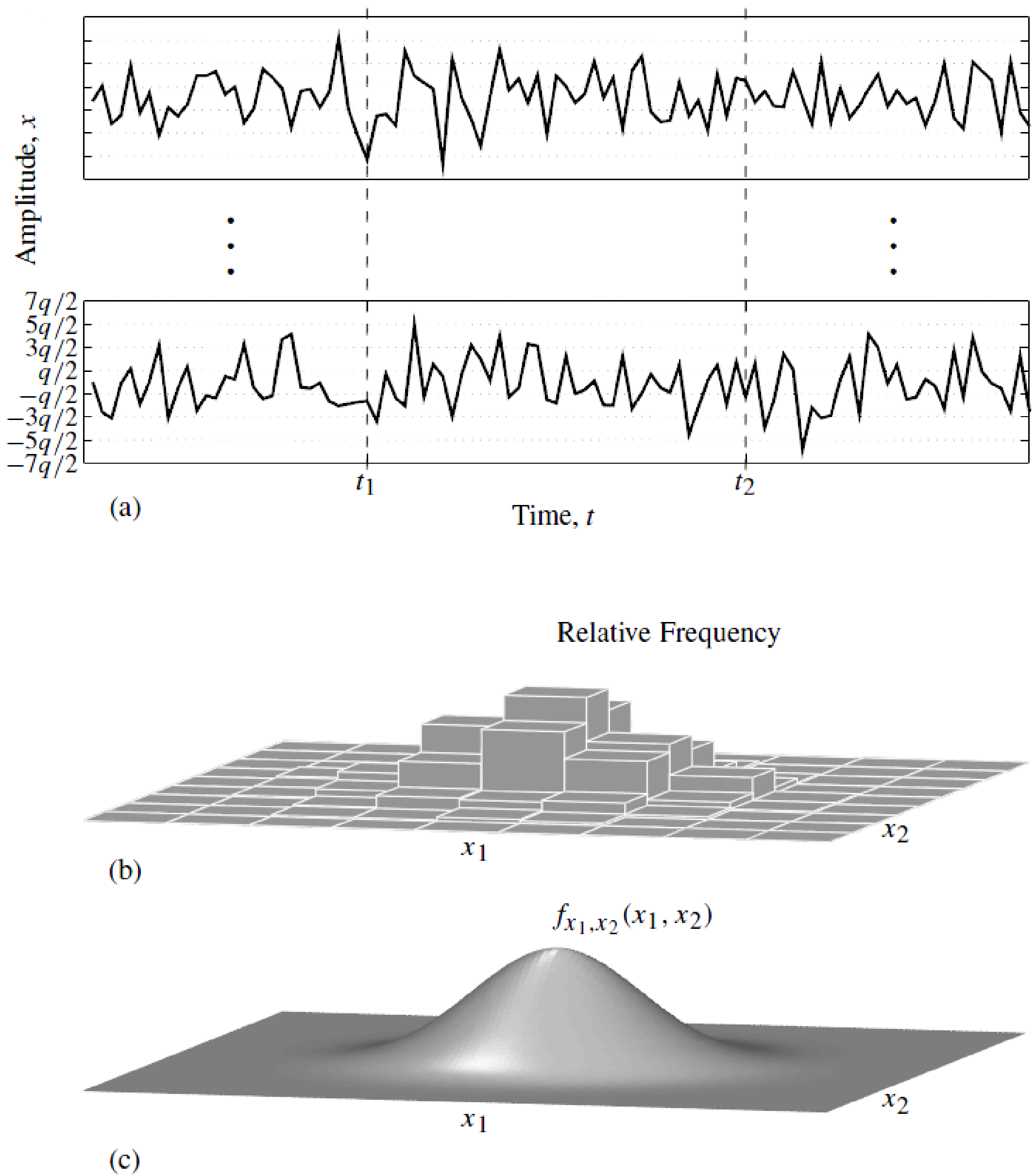


(b)

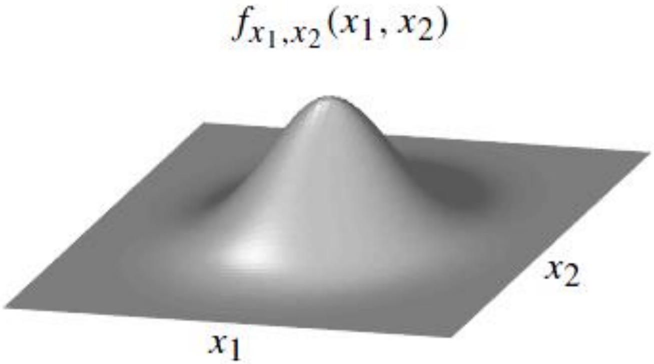
Derivation of a Histogram



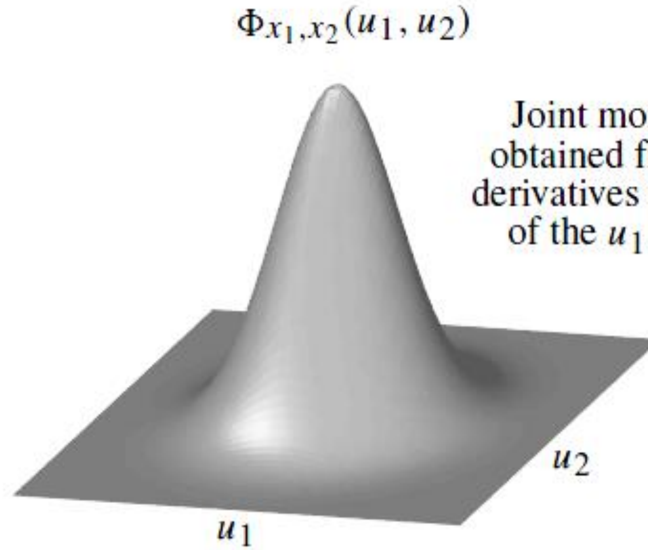
Derivation of a Two-Dimensional Histogram



PDF and CF



(a)



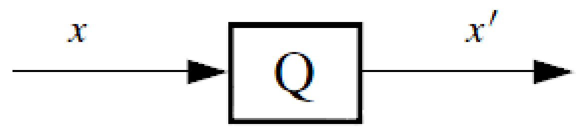
(b)

Joint moments are obtained from partial derivatives at the origin of the u_1, u_2 plane

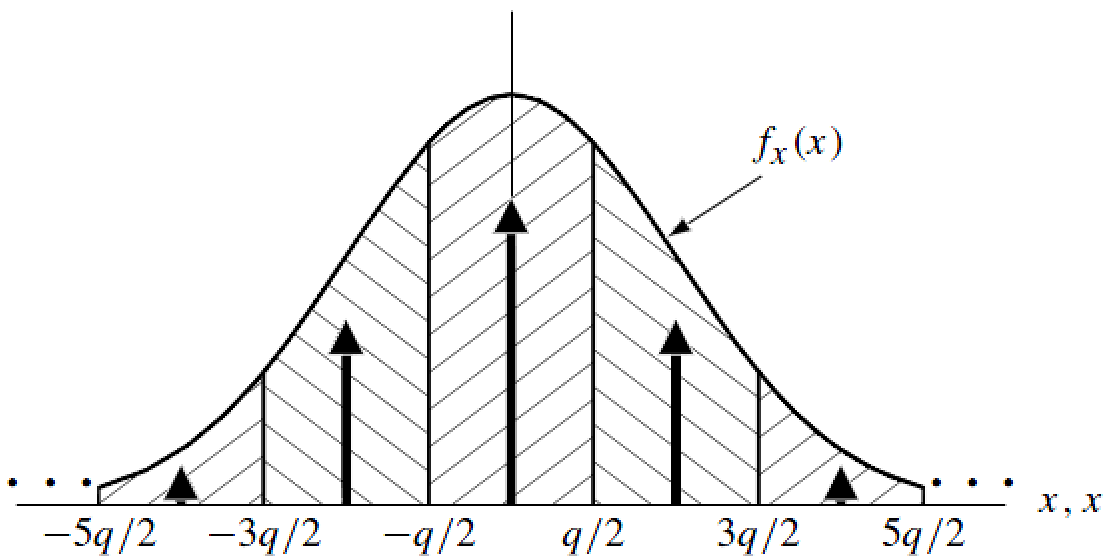
$$\Phi_{x_1,x_2}(u_1, u_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2) e^{j(u_1 x_1 + u_2 x_2)}\, \mathrm{d}x_1\, \mathrm{d}x_2$$
$$= \mathrm{E}\{e^{ju_1x_1 + ju_2x_2}\}.$$

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

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

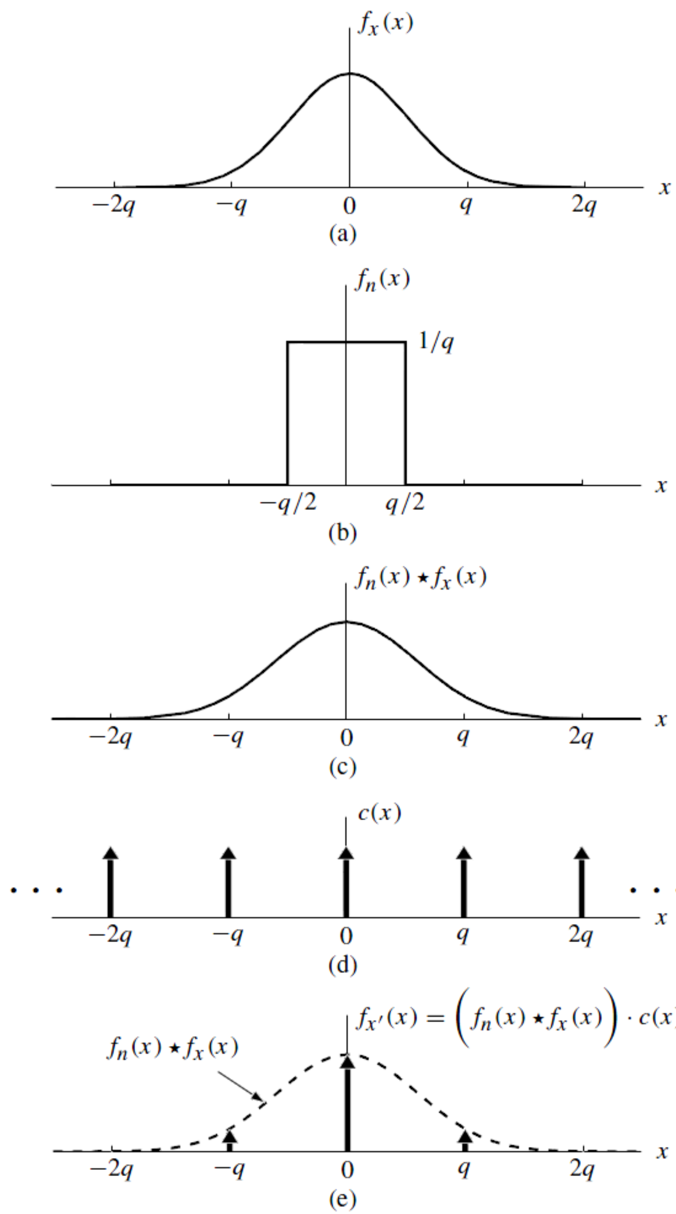


(a)

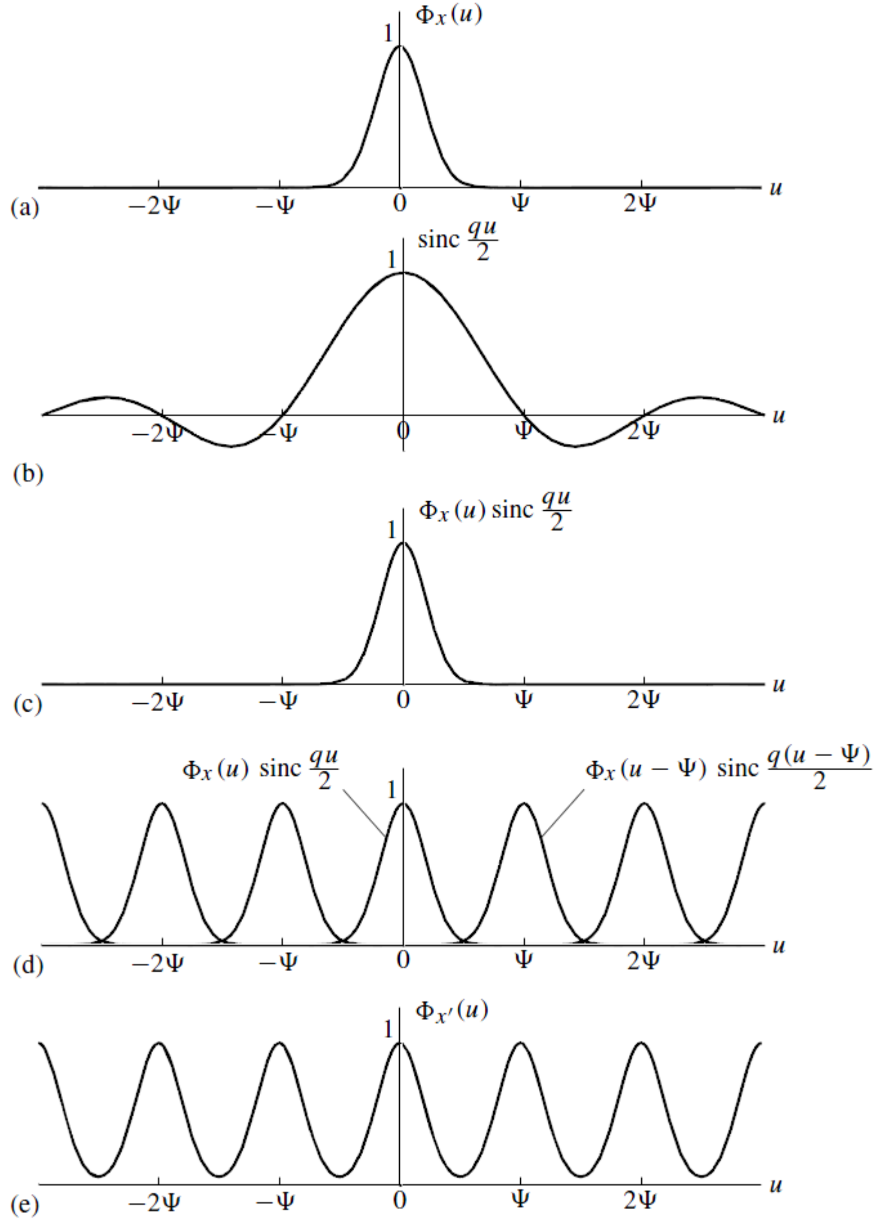


(b)

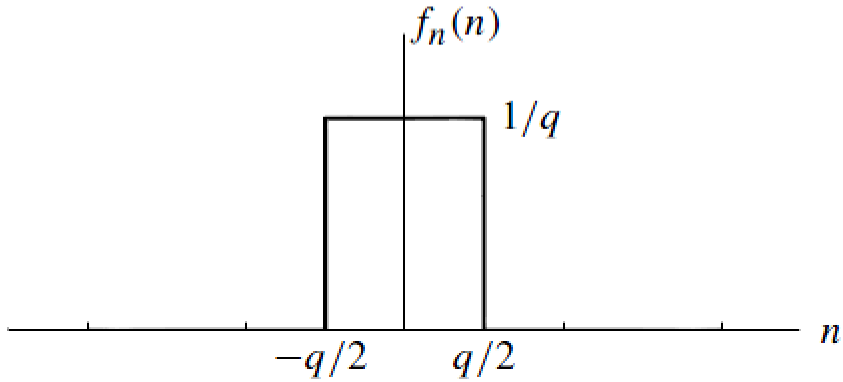
Derivation of PDF of x' from Area Sampling

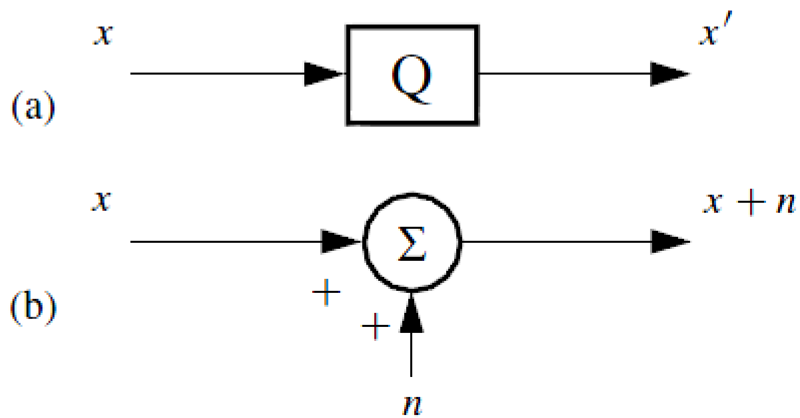


Area Sampling in the CF Domain

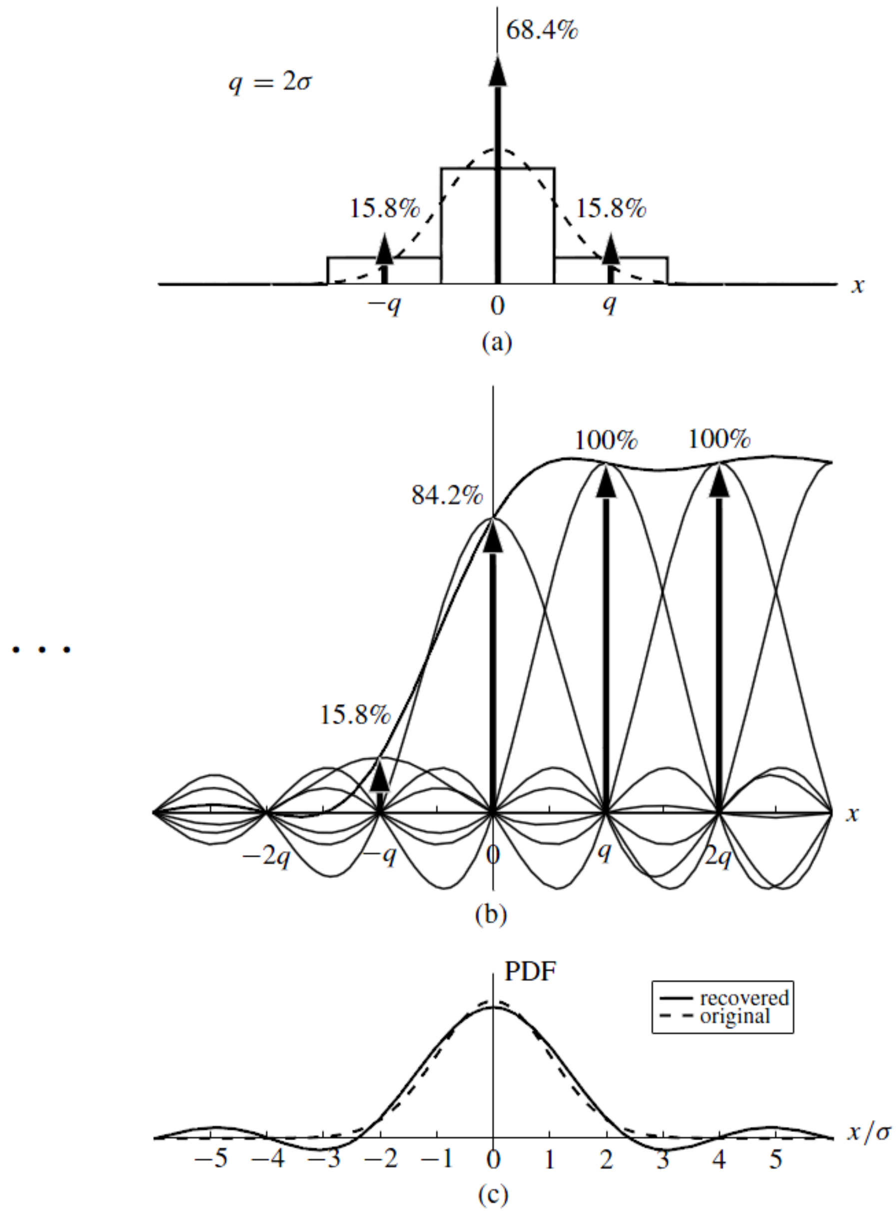


PQN Model

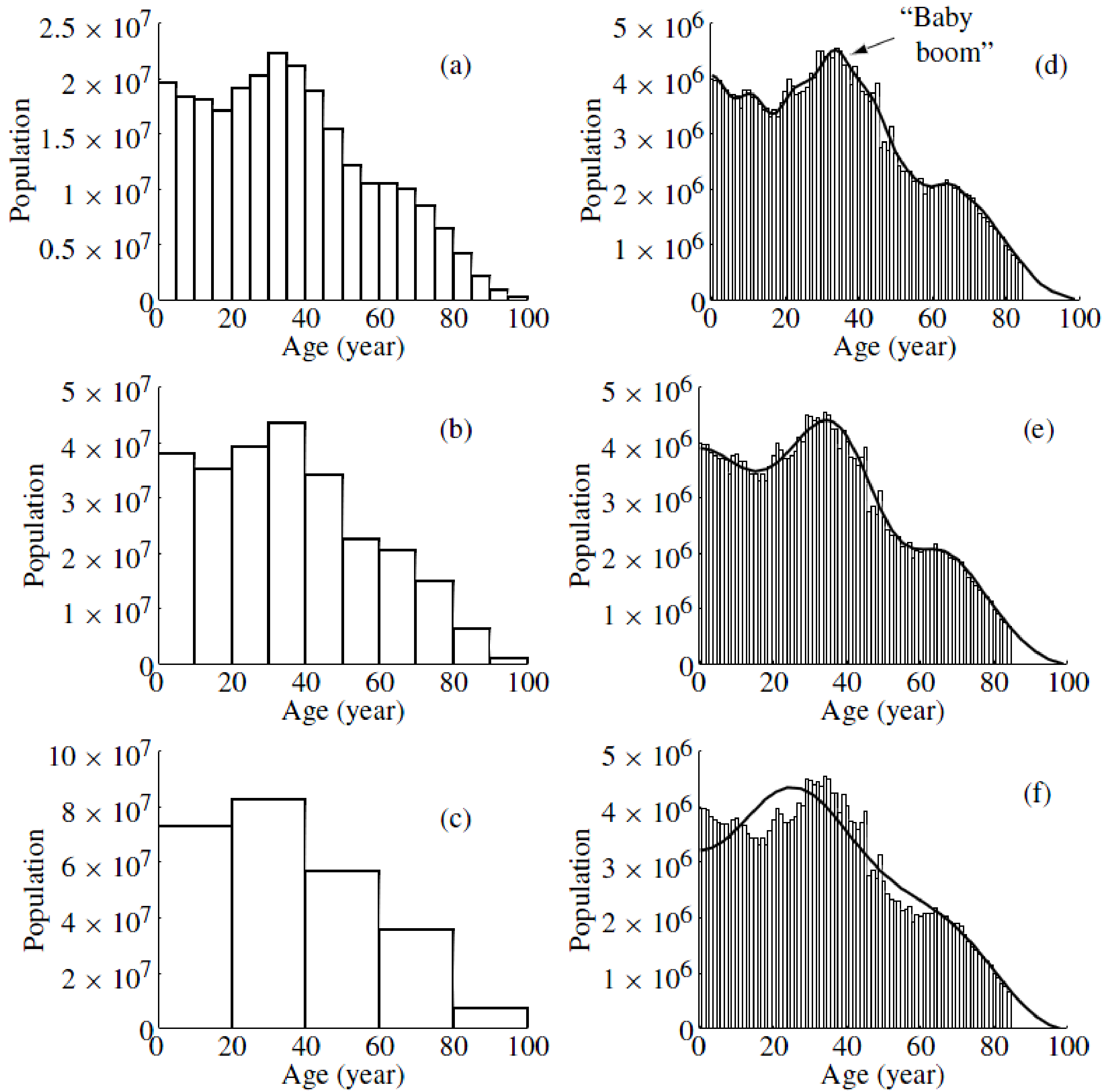




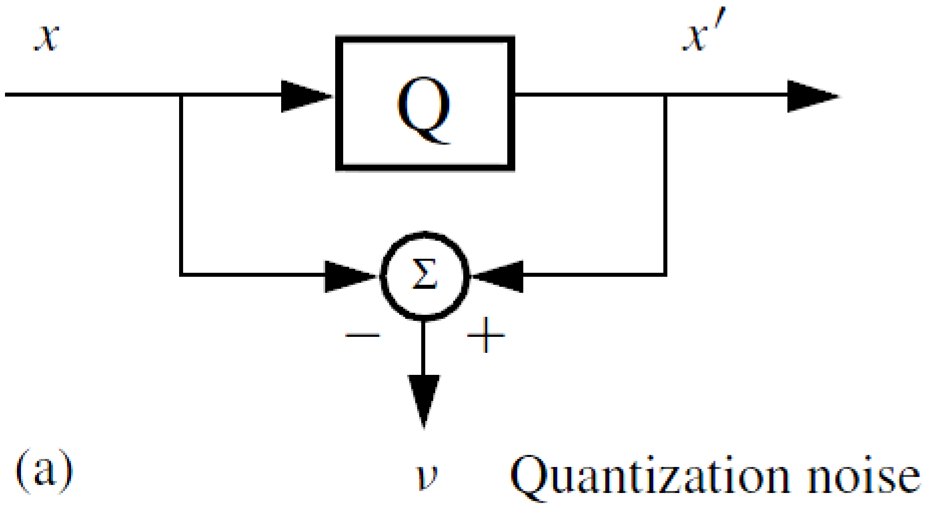
Original PDF from the Quantizer PDF



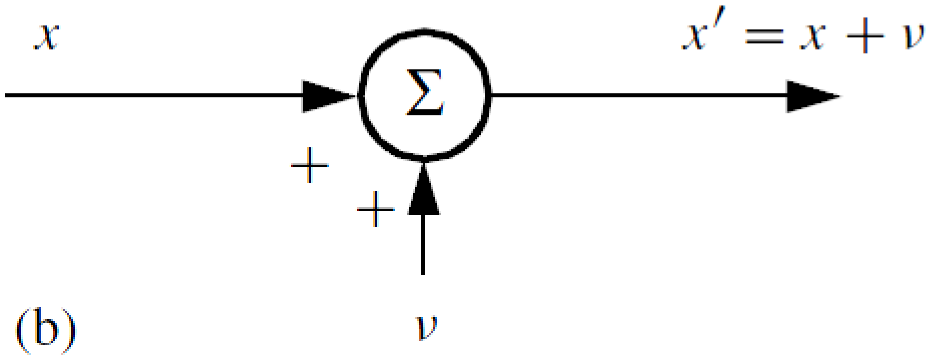
Reconstruction of 1990 U.S. Census Data



Quantization Noise

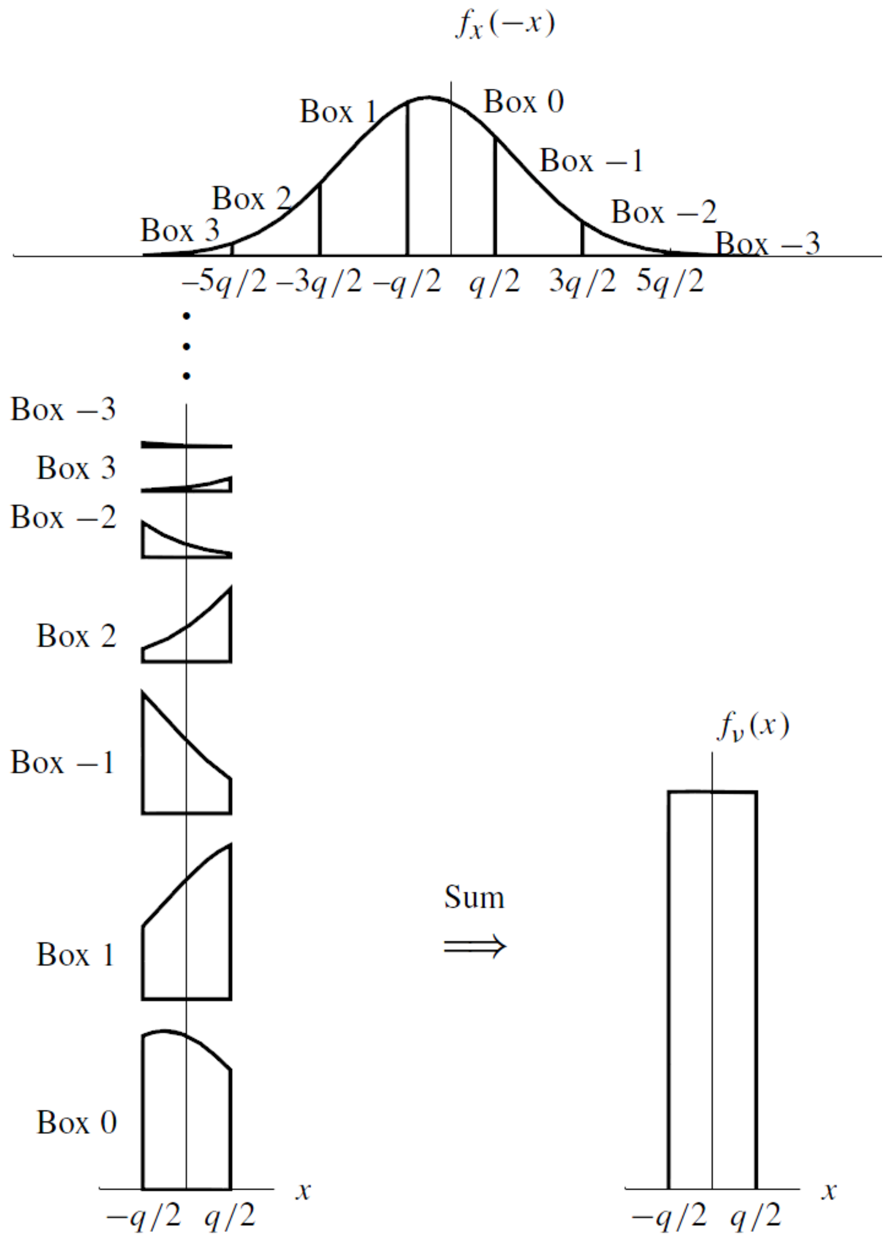


(a)

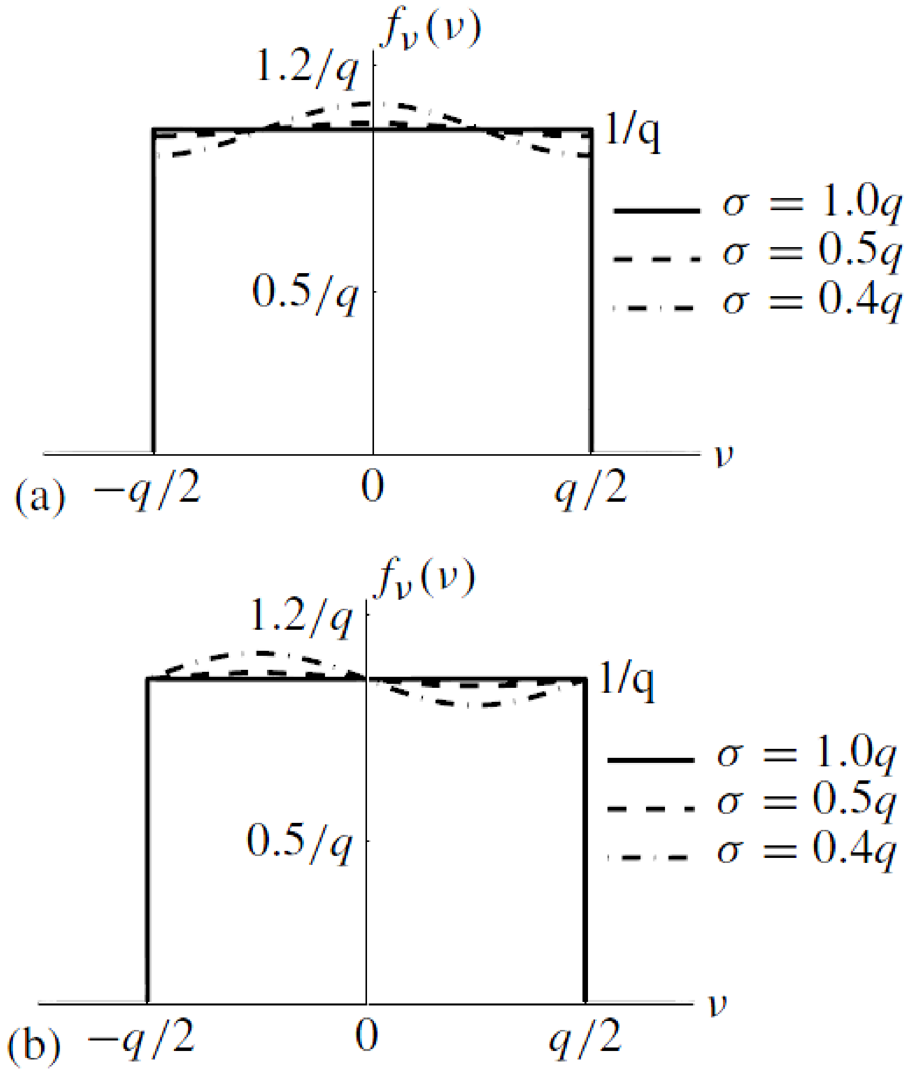


(b)

Construction of the PDF of Quantization Noise



PDF of Quantization Noise with Gaussian Input



Moments of Quantization Noise with Gaussian Input

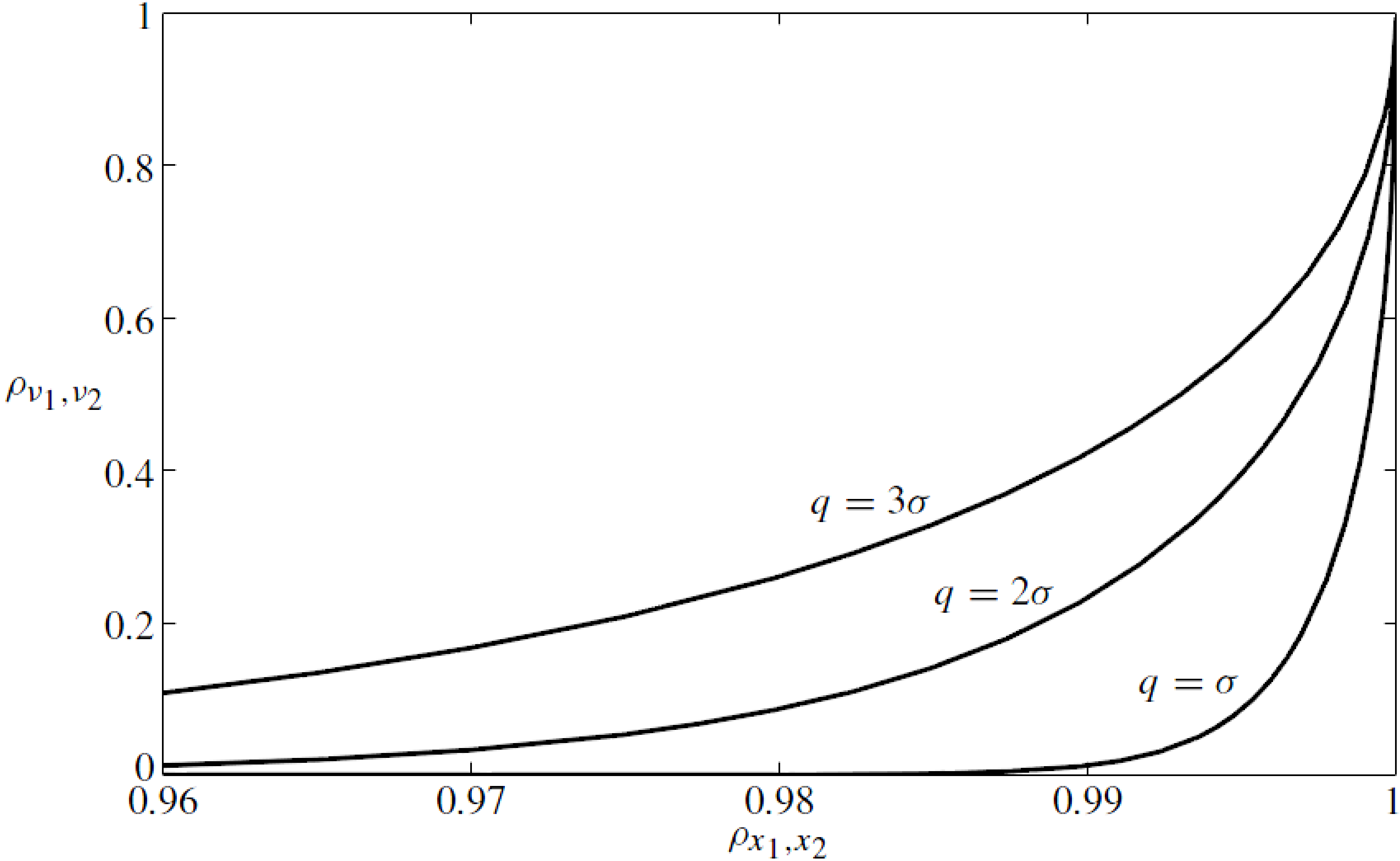
	$E\{\nu\}$	$E\{\nu^2\}$	$E\{\nu^3\}$	$E\{\nu^4\}$
$q = 2\sigma$	0	$\left(\frac{1}{12} - 7.3 \cdot 10^{-4}\right) q^2$	0	$\left(\frac{1}{80} - 3.6 \cdot 10^{-5}\right) q^4$
$q = 1.5\sigma$	0	$\left(\frac{1}{12} - 1.6 \cdot 10^{-5}\right) q^2$	0	$\left(\frac{1}{80} - 7.7 \cdot 10^{-7}\right) q^4$
$q = \sigma$	0	$\left(\frac{1}{12} - 2.7 \cdot 10^{-10}\right) q^2$	0	$\left(\frac{1}{80} - 1.3 \cdot 10^{-11}\right) q^4$
$q = 0.5\sigma$	0	$\left(\frac{1}{12} - 5.2 \cdot 10^{-36}\right) q^2$	0	$\left(\frac{1}{80} - 2.5 \cdot 10^{-37}\right) q^4$

	$E\{\nu\}$	$E\{\nu^2\}$	$E\{\nu^3\}$	$E\{\nu^4\}$
$q = 2\sigma$	$-2.3 \cdot 10^{-3} q$	$\left(\frac{1}{12} + 0\right) q^2$	$-2.2 \cdot 10^{-4} q^3$	$\left(\frac{1}{80} + 0\right) q^4$
$q = 1.5\sigma$	$-4.9 \cdot 10^{-5} q$	$\left(\frac{1}{12} + 0\right) q^2$	$-4.8 \cdot 10^{-6} q^3$	$\left(\frac{1}{80} + 0\right) q^4$
$q = \sigma$	$-8.5 \cdot 10^{-10} q$	$\left(\frac{1}{12} + 0\right) q^2$	$-8.3 \cdot 10^{-11} q^3$	$\left(\frac{1}{80} + 0\right) q^4$
$q = 0.5\sigma$	$-1.6 \cdot 10^{-35} q$	$\left(\frac{1}{12} + 0\right) q^2$	$-1.6 \cdot 10^{-36} q^3$	$\left(\frac{1}{80} + 0\right) q^4$

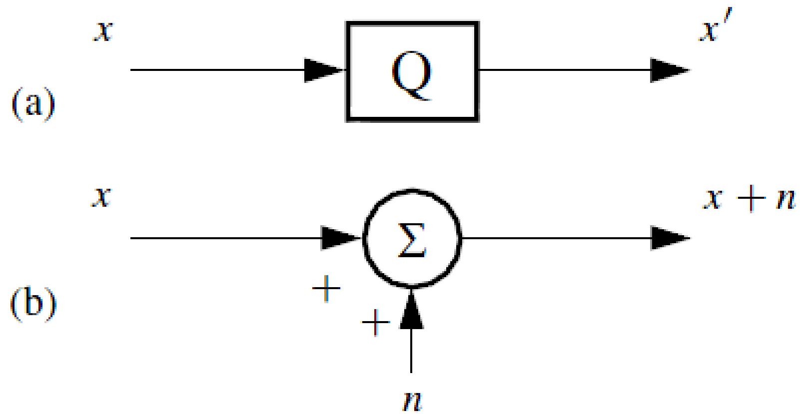
Correlation Coefficients between Gaussian Input and Noise

	$\frac{\mathrm{E}\{x\nu\}}{\sqrt{\mathrm{E}\{x^2\}\mathrm{E}\{\nu^2\}}} (= \rho_{x,\nu})$	$\mathrm{E}\{x'\nu\} = \mathrm{cov}\{x', \nu\}$	$\mathrm{E}\{xx'\} = \mathrm{cov}\{x, x'\}$
$q = 2\sigma$	$-2.50 \cdot 10^{-2}$	$\left(1 - 5.19 \cdot 10^{-2}\right) \frac{q^2}{12}$	$\left(1 - 1.44 \cdot 10^{-2}\right) \mathrm{E}\{x^2\}$
$q = 1.5\sigma$	$-7.15 \cdot 10^{-4}$	$\left(1 - 1.84 \cdot 10^{-3}\right) \frac{q^2}{12}$	$\left(1 - 3.10 \cdot 10^{-4}\right) \mathrm{E}\{x^2\}$
$q = \sigma$	$-1.85 \cdot 10^{-8}$	$\left(1 - 6.75 \cdot 10^{-8}\right) \frac{q^2}{12}$	$\left(1 - 5.35 \cdot 10^{-9}\right) \mathrm{E}\{x^2\}$
$q = 0.5\sigma$	$-7.10 \cdot 10^{-34}$	$\left(1 - 4.98 \cdot 10^{-33}\right) \frac{q^2}{12}$	$\left(1 - 1.02 \cdot 10^{-34}\right) \mathrm{E}\{x^2\}$

Relationship between ρ_{v_1,v_2} and ρ_{x_1,x_2}



Comparison of Quantization and PQN



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

This is the fundamental relation between the CFs of quantization and PQN. What Eq. (7.82) 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

$$\begin{aligned}
&\Phi_{x_1,\ldots,x_N,\nu_1,\ldots,\nu_N,x'_1,\ldots,x'_N}(u_{x_1},\ldots,u_{x_N},u_{\nu_1},\ldots,u_{\nu_N},u_{x'_1},\ldots,u_{x'_N}) \\
&\quad = \sum_{l_1=-\infty}^{\infty}\cdots\sum_{l_N=-\infty}^{\infty}\Phi_{x_1,\ldots,x_N,n_1,\ldots,n_N,x_1+n_1,\ldots,x_N+n_N}(u_{x_1},\ldots,u_{x_N}, \\
&\qquad\qquad u_{\nu_1},\ldots,u_{\nu_N},u_{x'_1}+l_1\Psi_1,\ldots,u_{x'_N}+l_N\Psi_N)\,.
\end{aligned}$$

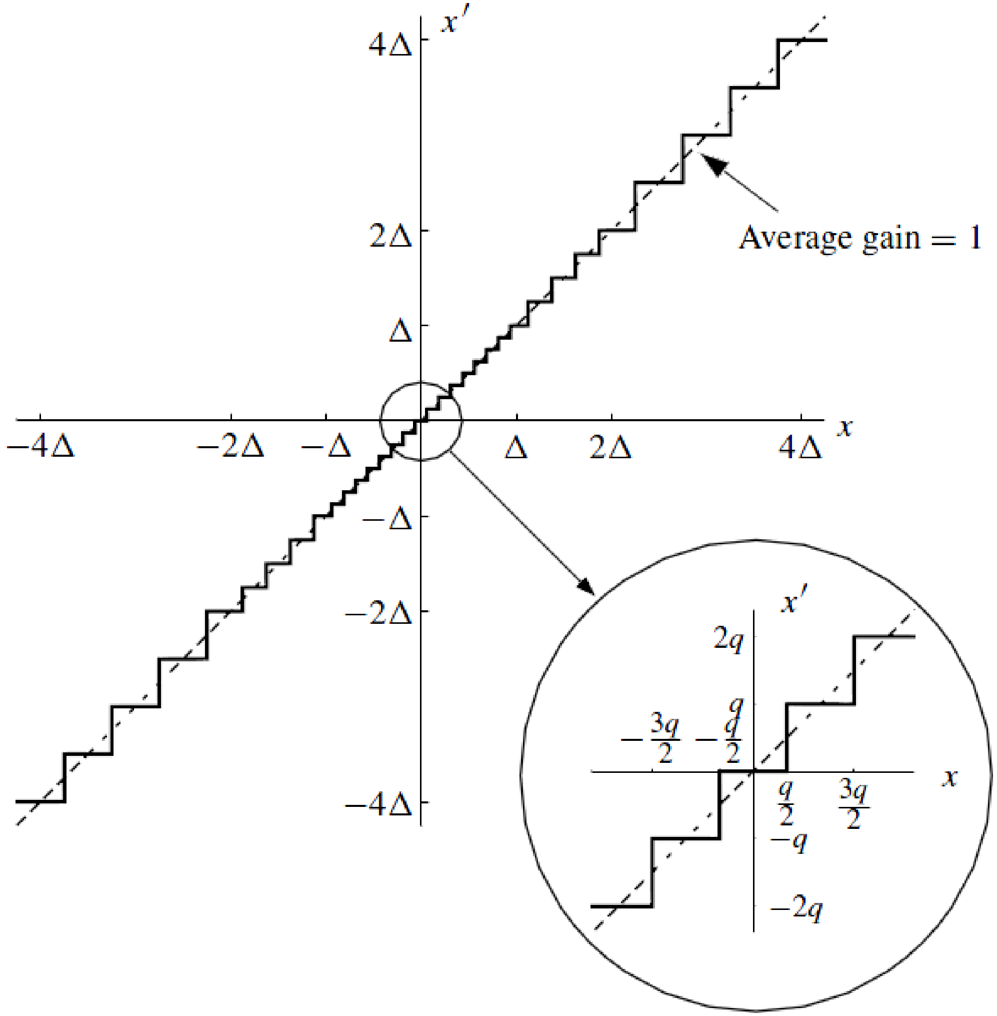
Uniform Quantization Summary

- If QT II is satisfied, the PQN model applies
- **Quantization of one variable:**
    - v is uniformly distributed
    - $E\{v\} = 0$
    - $E\{v^2\} = q^2/12$
    - $\mathrm{cov}\{x\ v\} = 0$
- **Quantization of two variables:**
    - All of the above applies to each of the variables
    - $E\{v_1\ v_2\} = 0$
- **Quantization of three or more variables:**
    - $E\{v_1 \ldots v_N\} = 0$
- If PQN applies, the quantizer may be replaced for purposes of analysis by a source of additive independent white noise.

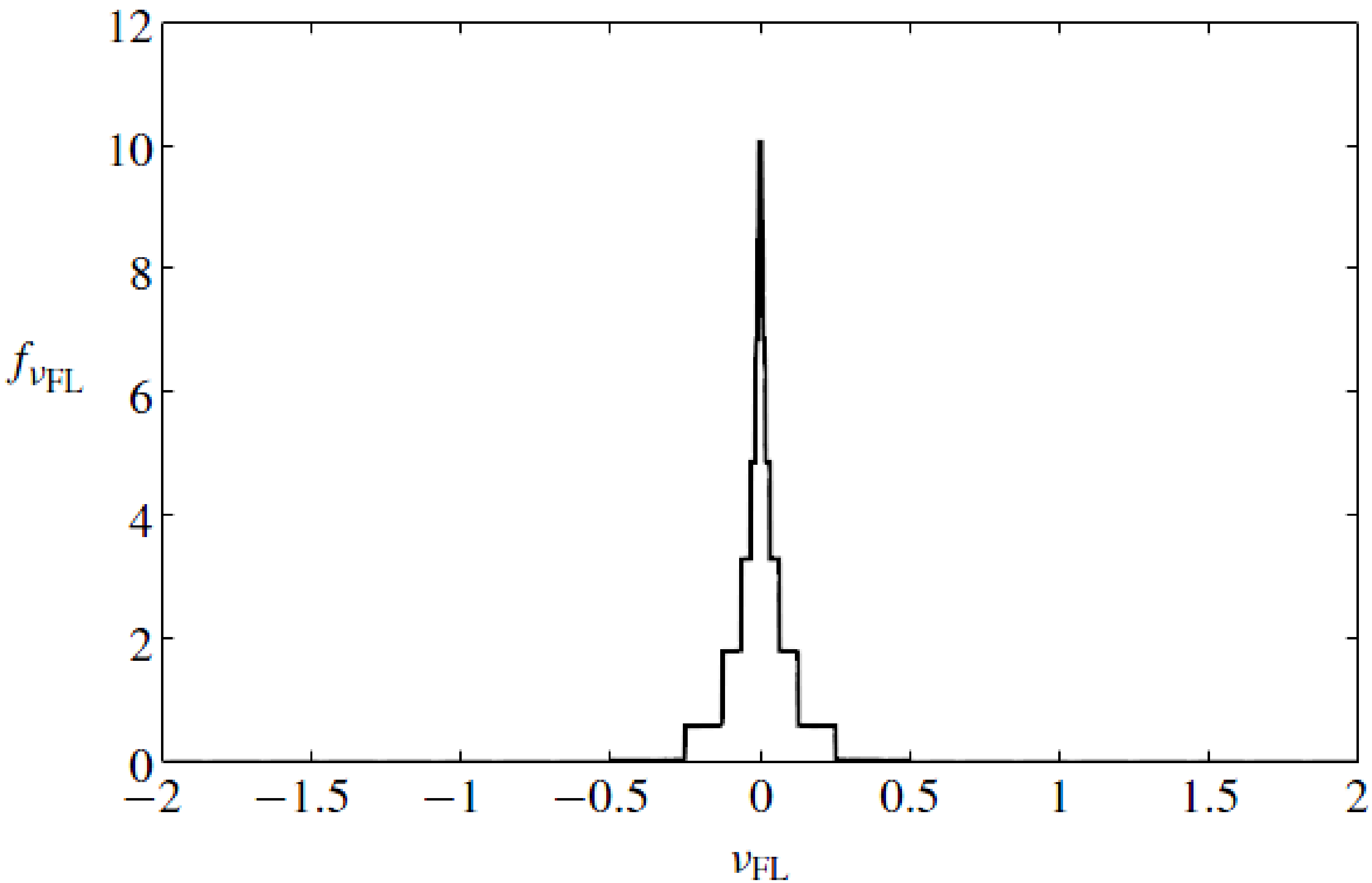
Counting with Binary Floating-Point Numbers with a 5-bit Mantissa

		Mantissa					
0		0	0	0	0	0	
1		0	0	0	0	1	
2		0	0	0	1	0	
3		0	0	0	1	1	
4		0	0	1	0	0	
5		0	0	1	0	1	
6		0	0	1	1	0	
7		0	0	1	1	1	
8		0	1	0	0	0	
9		0	1	0	0	1	
10		0	1	0	1	0	
11		0	1	0	1	1	
12		0	1	1	0	0	
13		0	1	1	0	1	
14		0	1	1	1	0	
15		0	1	1	1	1	
16	$\longrightarrow$	1	0	0	0	0	$\times 2^E$
17		1	0	0	0	1	
18		1	0	0	1	0	
19		1	0	0	1	1	
20		1	0	1	0	0	
21		1	0	1	0	1	
22		1	0	1	1	0	
23		1	0	1	1	1	
24		1	1	0	0	0	
25		1	1	0	0	1	
26		1	1	0	1	0	
27		1	1	0	1	1	
28		1	1	1	0	0	
29		1	1	1	0	1	
30		1	1	1	1	0	
31	$\longleftarrow$	1	1	1	1	1	

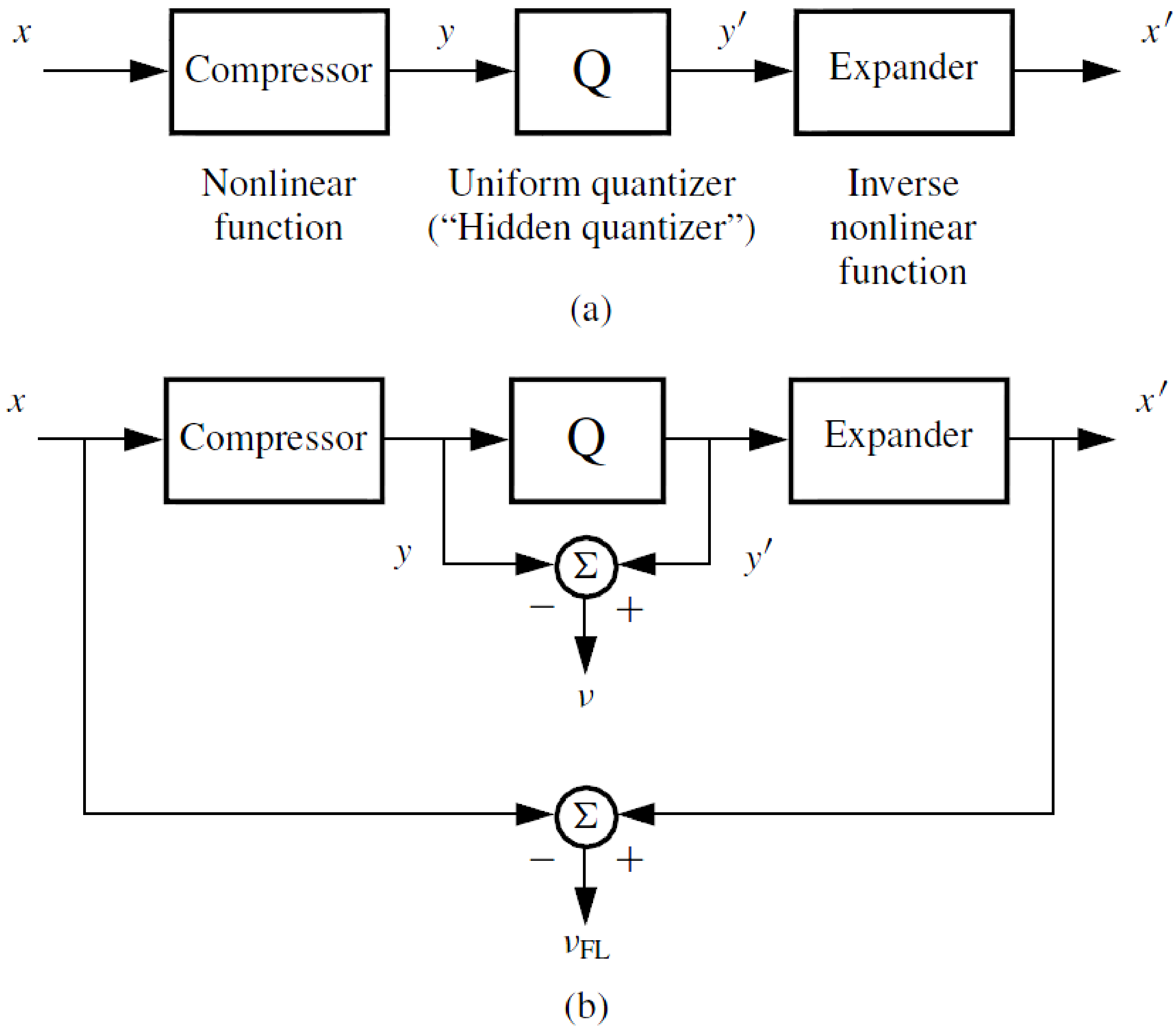
A Floating-Point Quantizer with a 3-bit Mantissa



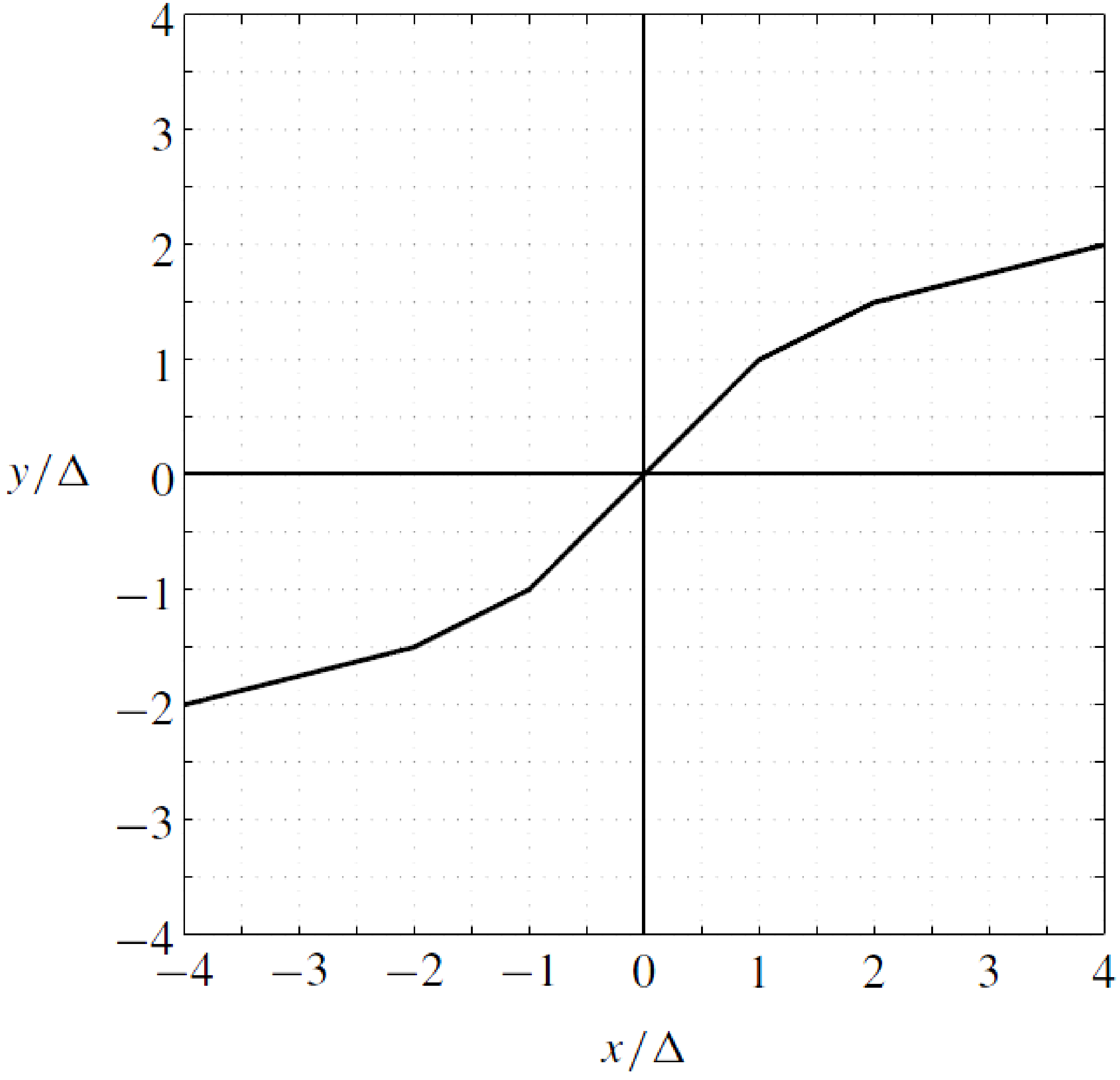
PDF of Floating-Point Quantization Noise



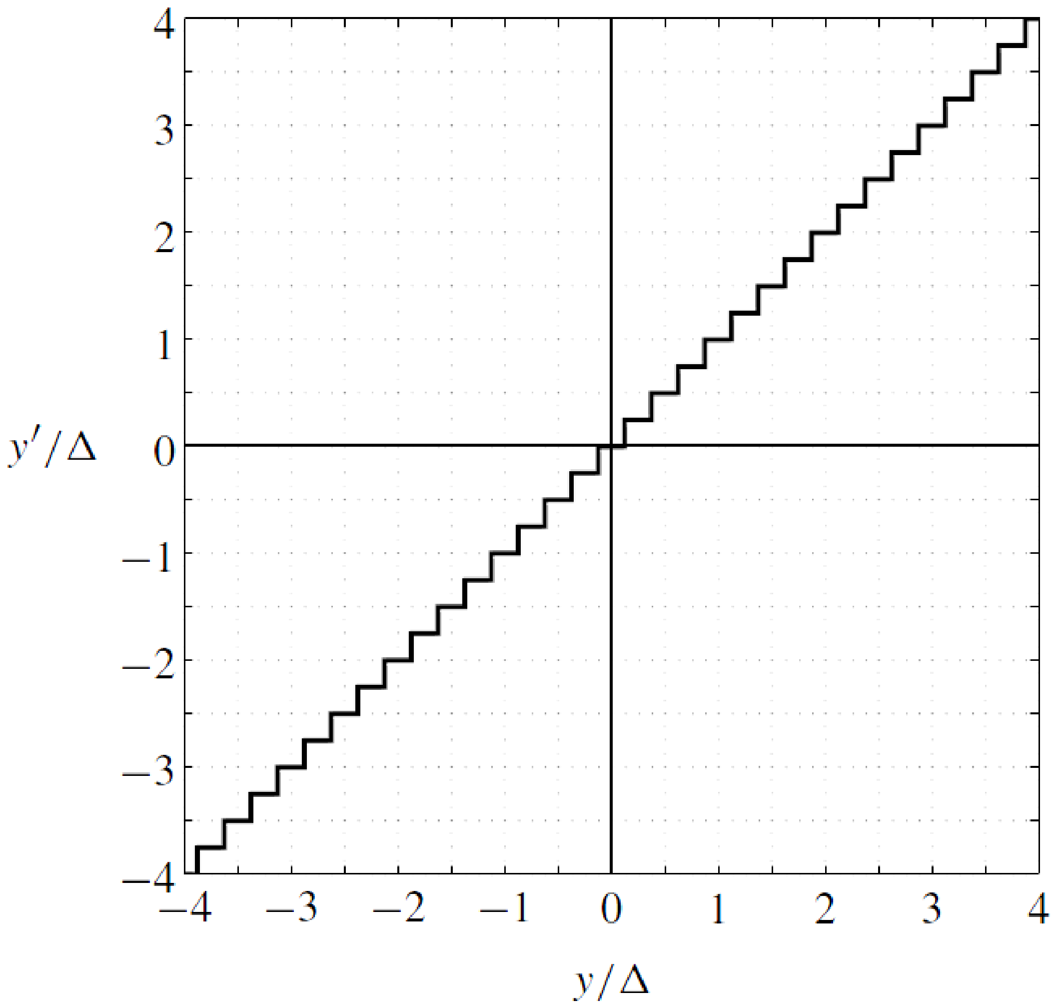
A Model of a Floating-Point Quantizer



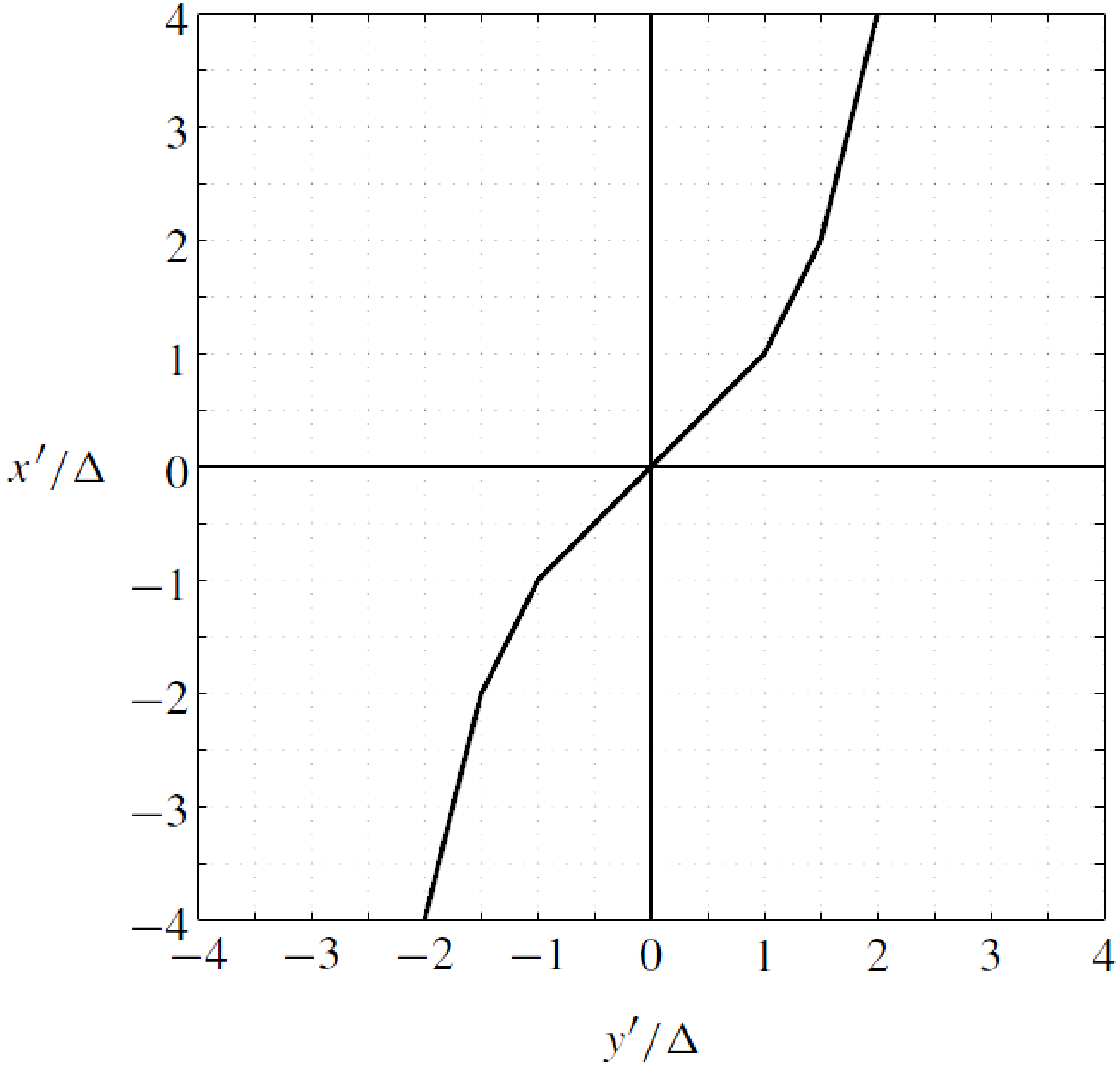
The Compressor's Input-Output Characteristic



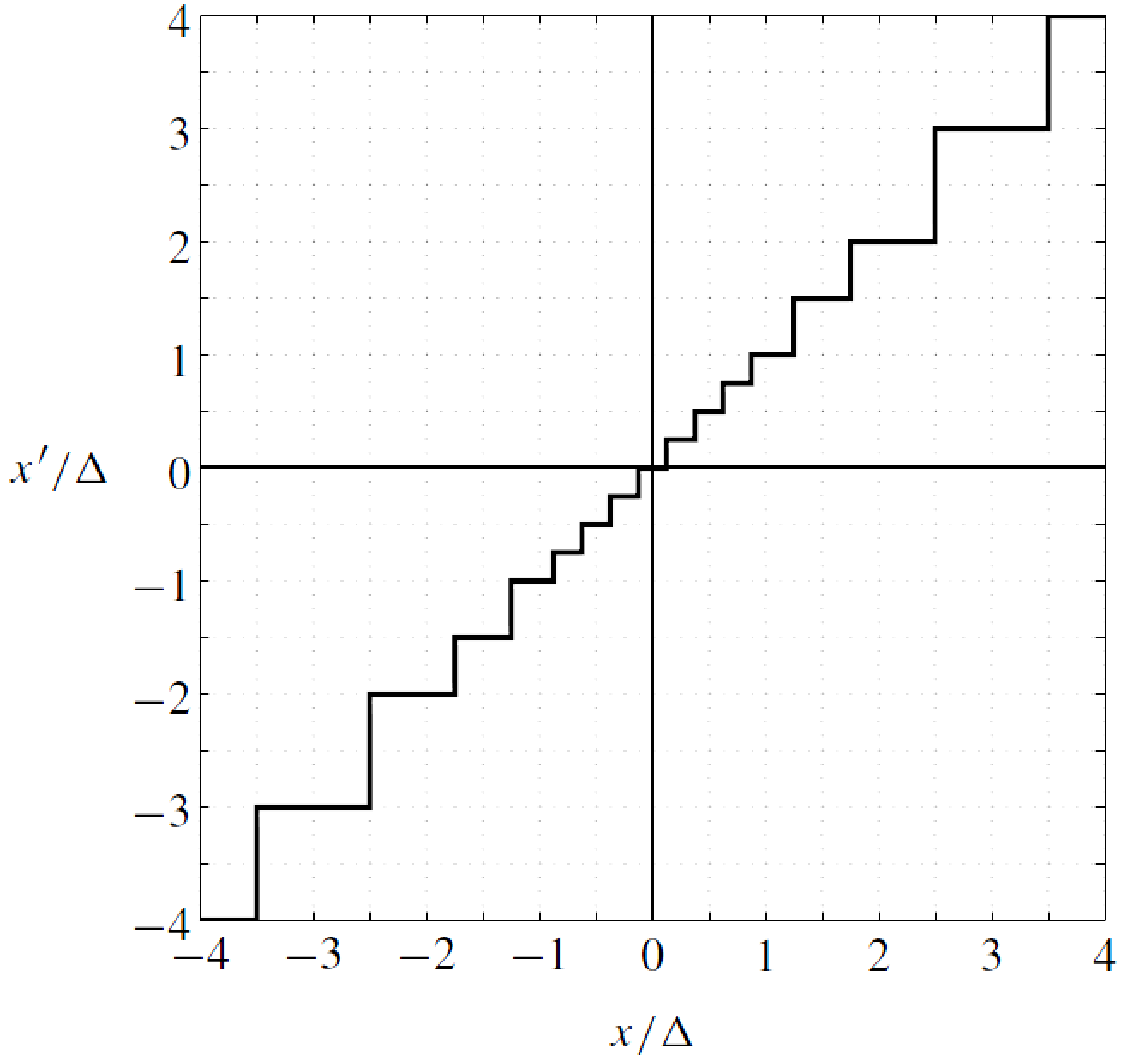
The Uniform "Hidden Quantizer" with a 2-bit Mantissa



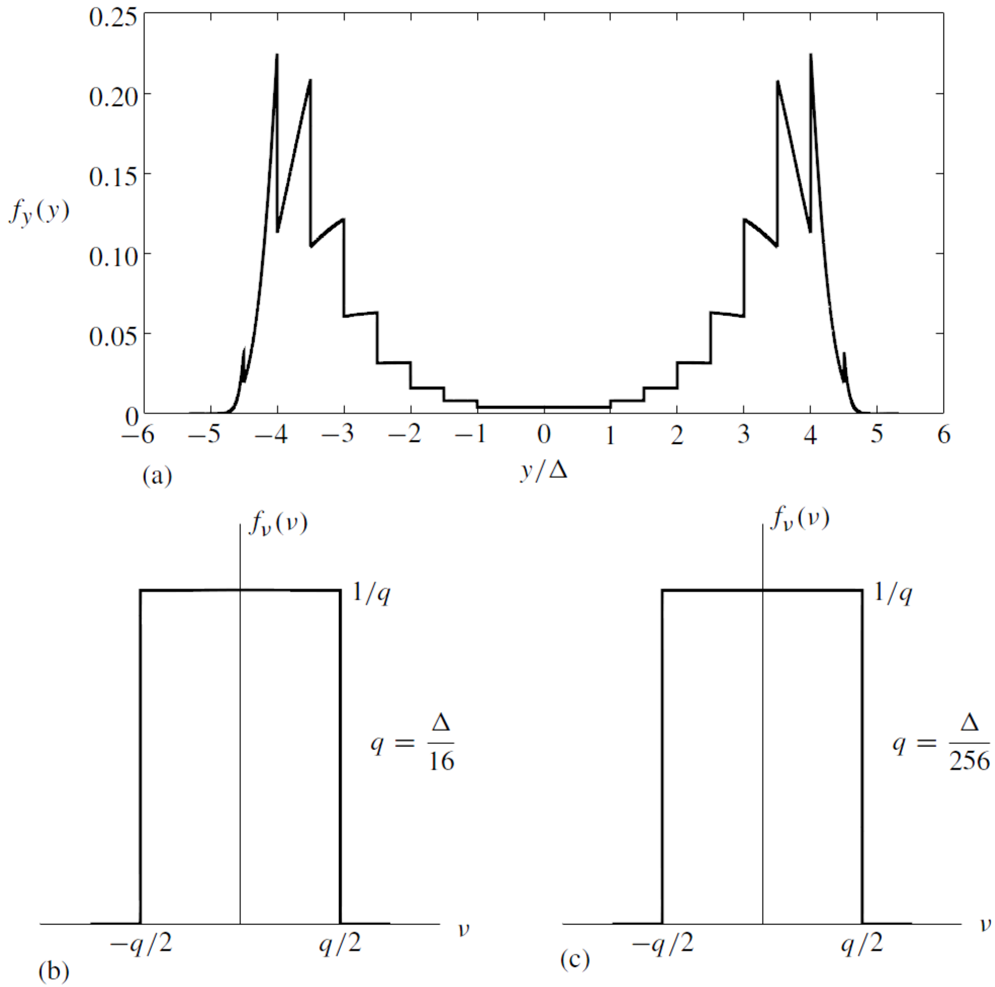
The Expandor's Input-Output Characteristic



A Floating-Point Quantizer with a 2-bit Mantissa



Compressor Output and Hidden Quantization Noise



PDF of compressor output and of hidden quantization noise when x is zero-mean Gaussian with $\sigma_x = 50\Delta$: (a) $f_y(y)$; (b) $f_\nu(\nu)$ for $p = 4$ $(q = \Delta/16)$; (c) $f_\nu(\nu)$ for $p = 8$ $(q = \Delta/256)$.

Floating-Point Quantization Summary

- When the PQN model applies to the "hidden quantizer":
  - **Quantization of one variable:**

$$\nu_{FL} \text{ is "skyscraper-distributed"}$$
$$\mathrm{E}\{\nu_{FL}\} = 0$$
$$\left(\frac{1}{12}\right)2^{-2p}\,\mathrm{E}\{x^2\} \le \mathrm{E}\{\nu_{FL}^2\} \le \left(\frac{1}{3}\right)2^{-2p}\,\mathrm{E}\{x^2\}$$
$$\mathrm{E}\{\nu_{FL}^2\} \approx 0.180 \times 2^{-2p}\,\mathrm{E}\{x^2\}$$
$$SNR = \frac{\mathrm{E}\{x^2\}}{\mathrm{E}\{\nu_{FL}^2\}}$$
$$3 \times 2^{2p} \le SNR \le 12 \times 2^{2p}\ ; \quad SNR \approx 5.55 \times 2^{2p}$$
$$\text{With a 6-bit mantissa, } SNR \approx 5.55 \times 2^{2p} = 2.38 \times 10^{10} \ \ (104\,\mathrm{dB})$$
$$\mathrm{cov}\{x\,\nu_{FL}\} = 0$$

  - **Quantization of multiple variables:**

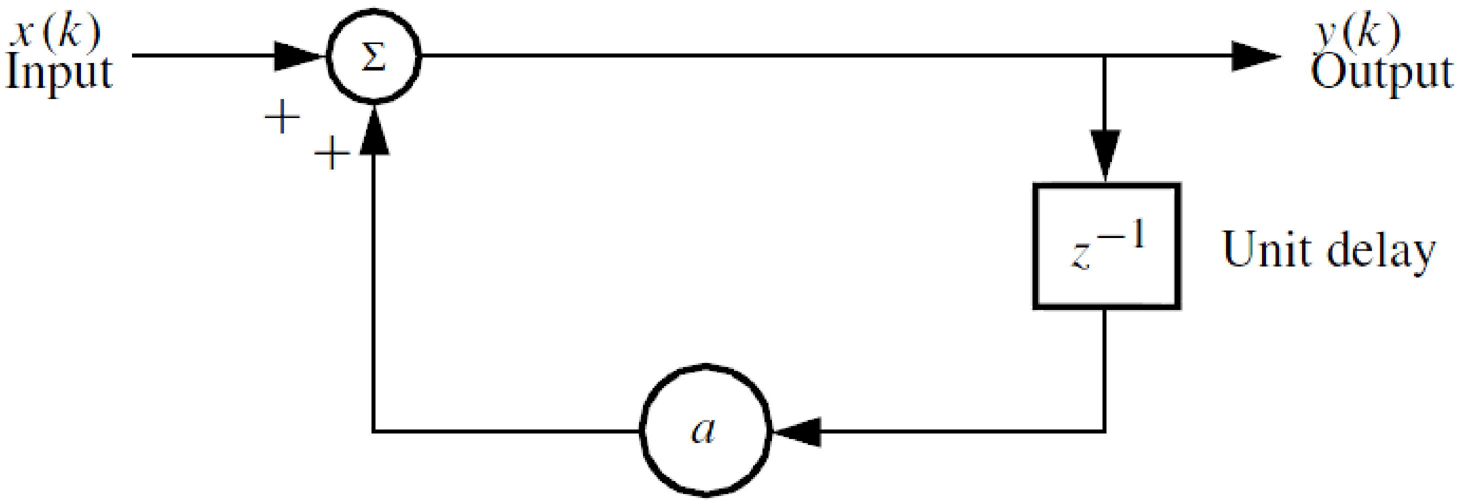
$$\text{All of the above applies to each of the variables}$$
$$\mathrm{E}\{\nu_{FL_1}\nu_{FL_2} \cdots \nu_{FL_N}\} = 0$$

PQN Model for "Hidden Quantizer" Works!

	Mean of noise	**Normalized mean square of quantization noise:**	**Correlation coefficient of v_{FL1} and x**	**Correlation coefficient of v_{FL1} and v_{FL2}**	
				$\rho_{x1,x2}$ = 0.99	$\rho_{x1,x2}$ = 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.

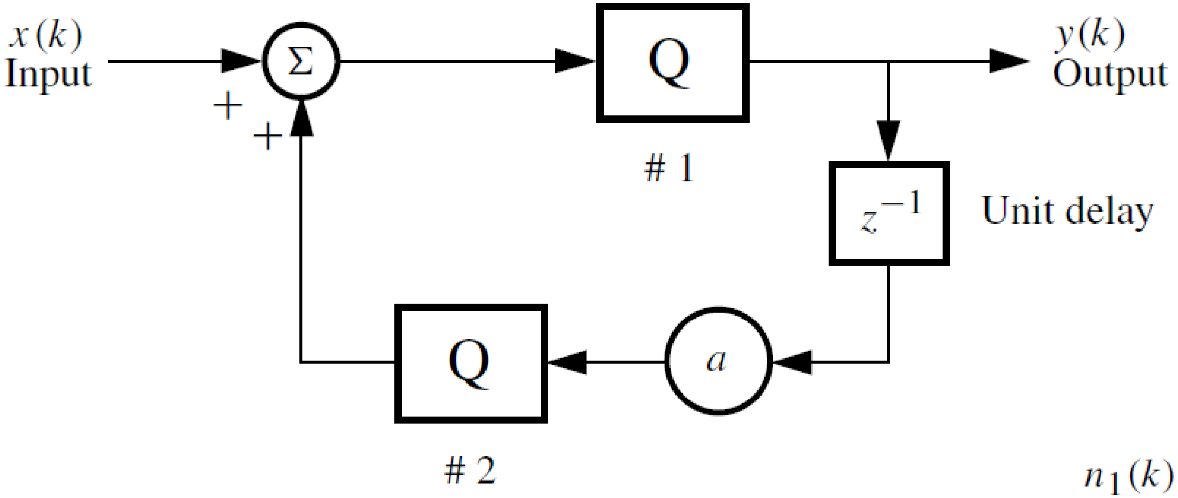
Roundoff in Digital Filters

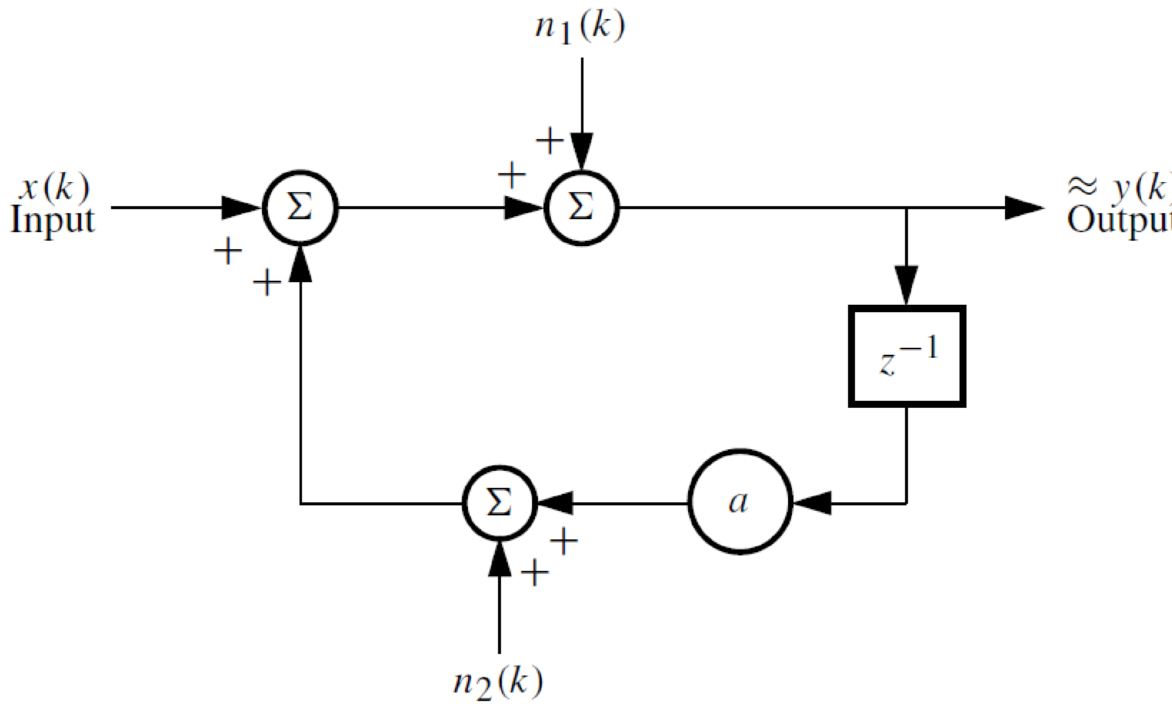


$$y(k) - ay(k-1) = x(k)$$

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

Implementation and Analysis





Quantizer Noise Power

$$\begin{pmatrix} \text{transfer function} \\ \text{from quantizer \#1} \\ \text{to filter output} \end{pmatrix} = \frac{1}{1 - az^{-1}} \qquad \begin{pmatrix} \text{impulse response} \\ \text{from quantizer \#1} \\ \text{to filter output} \end{pmatrix} = 1, a, a^2, a^3, \ldots$$

$$\begin{pmatrix} \text{sum of} \\ \text{of} \\ \text{squares} \end{pmatrix} = 1 + a^2 + a^4 \cdots = \frac{1}{1 - a^2}$$

Assuming a white, Gaussian, zero-mean input, the noise power of each quantizer is:

$$\frac{q^2}{12} \cdot \begin{pmatrix} \text{sum} \\ \text{of} \\ \text{squares} \end{pmatrix} = \frac{q^2/12}{1 - a^2}$$

Since both quantizers generate the same noise power, the total noise at the output is:

$$\begin{pmatrix} \text{total} \\ \text{output} \\ \text{quantization} \\ \text{noise power} \end{pmatrix} = \frac{q^2/6}{1 - a^2}$$

Signal-to-Noise Ratio (SNR)

$$
\begin{pmatrix} \text{output} \\ \text{SNR} \end{pmatrix} = \frac{\frac{1}{2}(\frac{A}{1-a})^2}{(\frac{q^2/6}{1-a^2})}
$$

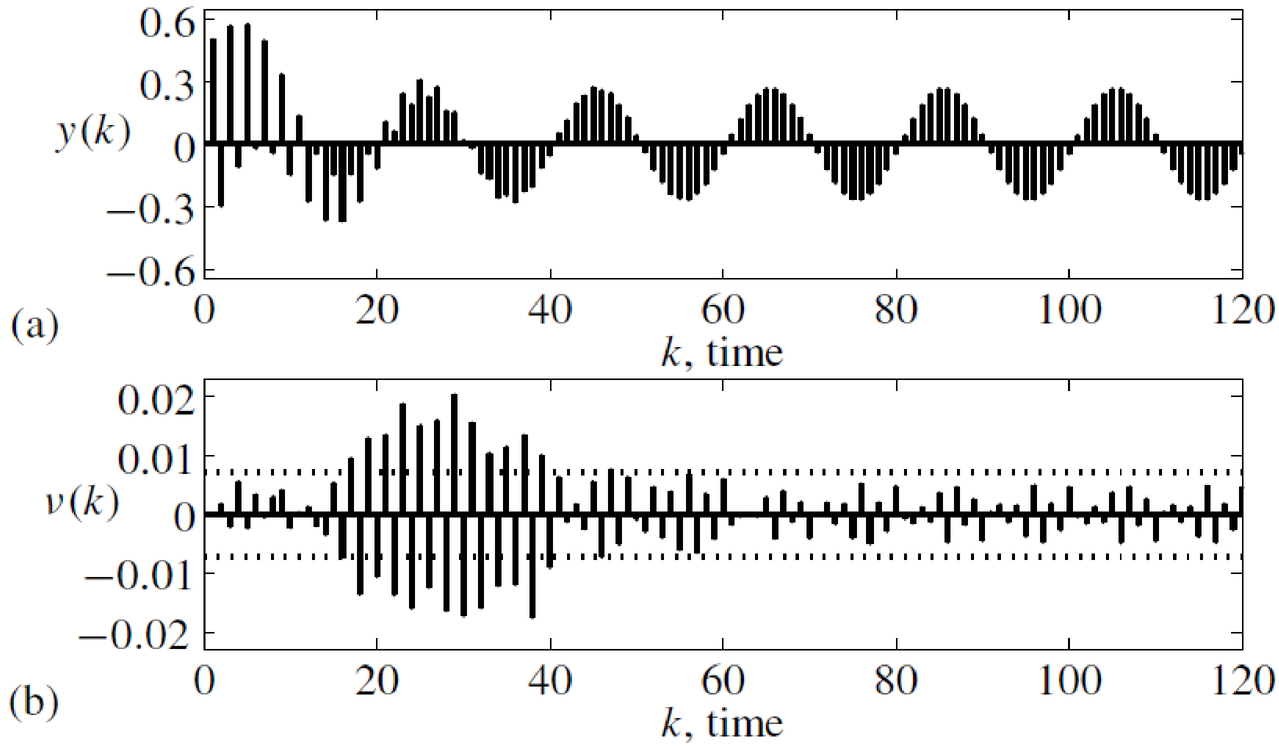
$$
= \frac{\frac{1}{2}q^2 2^{22}}{(\frac{q^2/6}{1-a^2})}
$$

$$
= (1-a^2)3 \cdot 2^{22}
$$

$$
= 1.26 \cdot 10^7(1-a^2)
$$

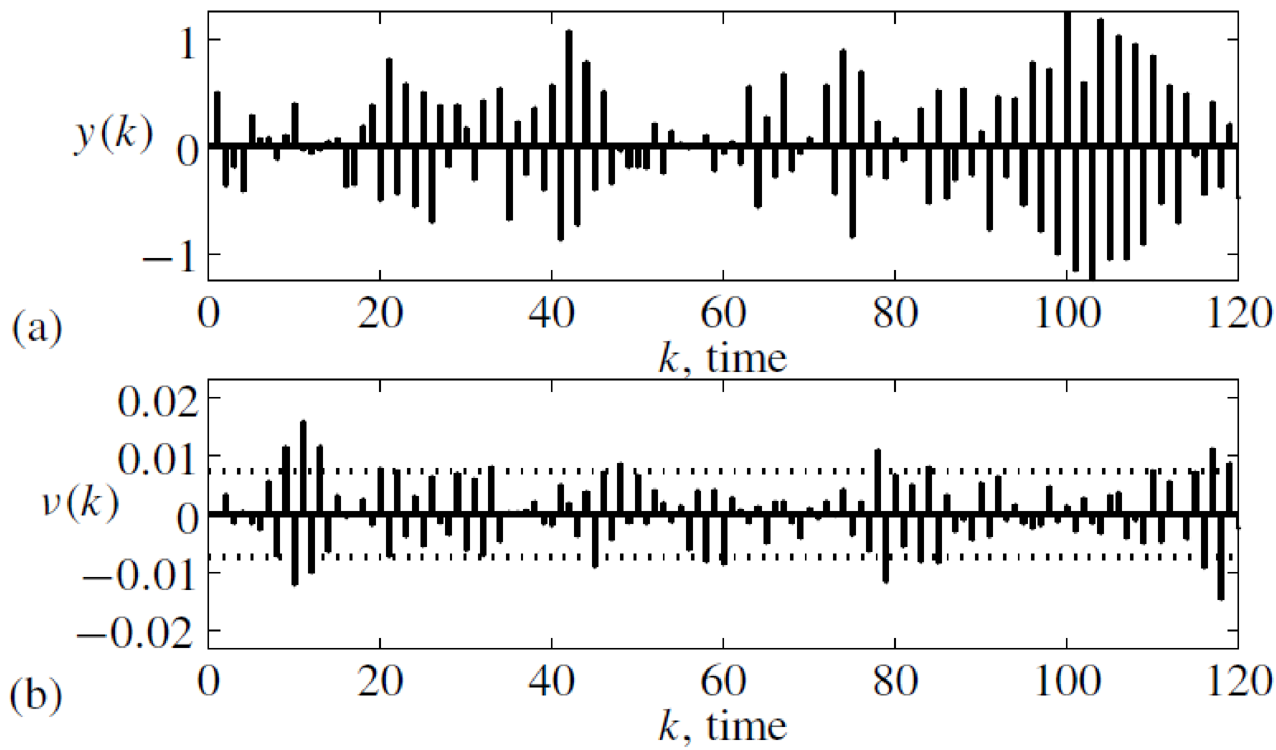
$$
= 71.0 + 10\log_{10}(1-a^2)\ \text{dB}
$$

Example with Sine Wave Input

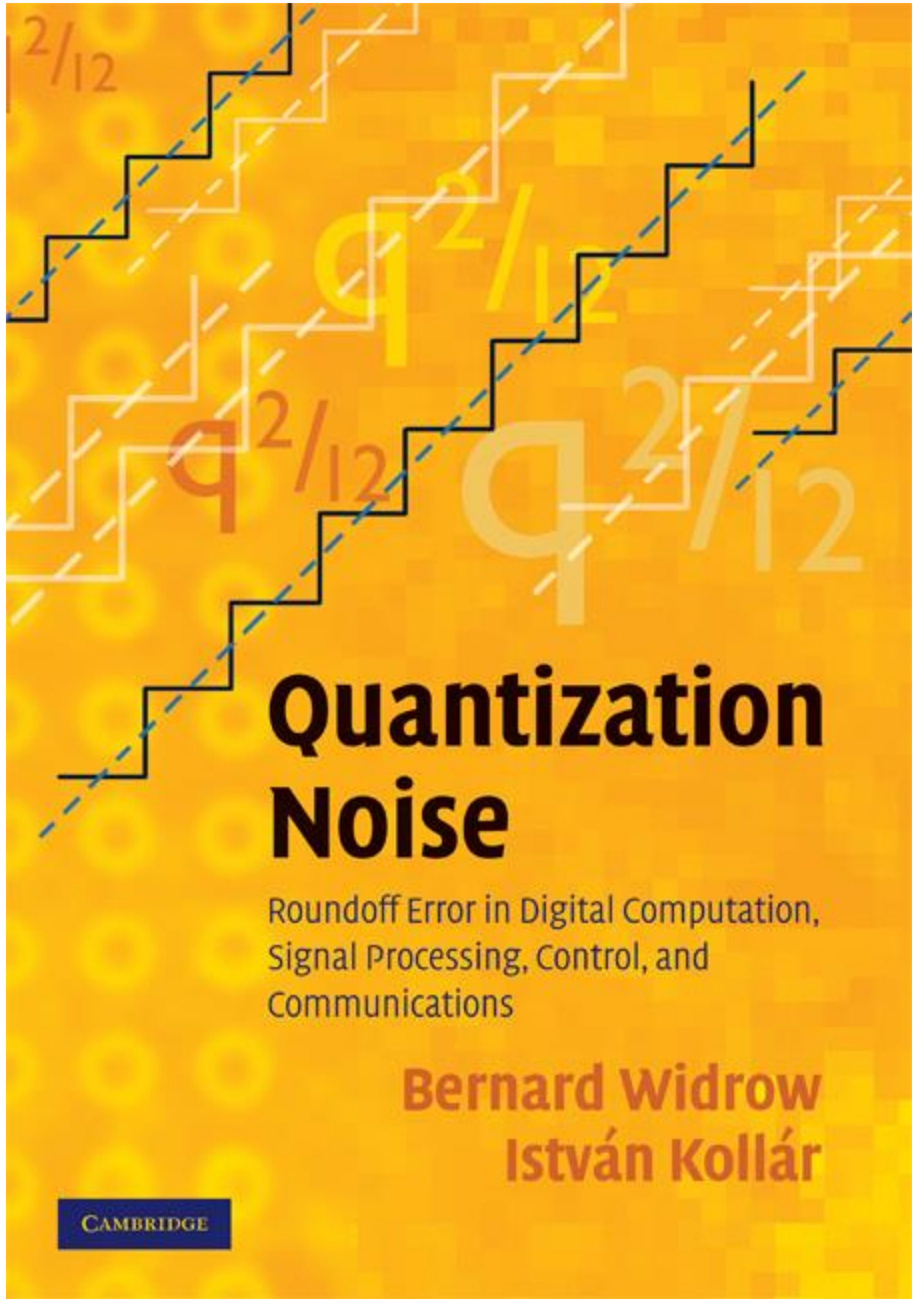


Response to a sine wave with frequency 1/20 the sample rate applied to the one-pole filter on the previous slides. Input begins at k = 0; quantizers are 8-bit working on the amplitude range [-1, 1]; parameter a = 0.11101 in binary (approximately 0.906). (a) output response; (b) output quantization noise, with theoretical standard deviations marked with dotted lines.

Example with Gaussian Input



Response to a white-noise zero-mean Gaussian input with σ = 0.25 applied to the one-pole filter on the previous slides. Input begins at k = 0; quantizers are 8-bit working on the amplitude range [-1, 1]; parameter a = 0.11101 in binary (approximately 0.906). (a) output response; (b) output quantization noise, with theoretical standard deviations marked with dotted lines.


q²/12
Quantization
Noise
Roundoff Error in Digital Computation, Signal Processing, Control, and Communications
Bernard Widrow
István Kollár
CAMBRIDGE