Quantization Noise

Roundoff Error in Digital Computation, Signal Processing, Control, and Communications

Bernard Widrow
István Kollár

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521886710

This page intentionally left blank

Quantization Noise

If you are working in digital signal processing, control, or numerical analysis, you will find this authoritative treatment of quantization noise (roundoff error) to be an invaluable resource.

Do you know where the theory of quantization noise comes from, and under what circumstances it is true? Expert authors, including the founder of the field and formulator of the theory of quantization noise, Bernard Widrow, answer these and other important practical questions. They describe and analyze uniform quantization, floating-point quantization, and their applications in detail.

Key features include:

- heuristic explanations along with rigorous proofs;
- worked examples, so that theory is understood through examples;
- focus on practical cases, engineering approach;
- analysis of floating-point roundoff;
- dither techniques and implementation issues analyzed;
- program package for MATLAB® available on the web, for simulation and analysis of fixed-point and floating-point roundoff;
- homework problems and solutions manual; and
- actively maintained website with additional text on special topics on quantization noise.

The additional resources are available online through www.cambridge.org/9780521886710

BERNARD WIDROW, an internationally recognized authority in the field of quantization, is a Professor of Electrical Engineering at Stanford University, California. He pioneered the field and one of his papers on the topic is the standard reference. He is a Fellow of the IEEE and the AAAS, a member of the US National Academy of Engineering, and the winner of numerous prestigious awards.

ISTVÁN KOLLÁR is a Professor of Electrical Engineering at the Budapest University of Technology and Economics. A Fellow of the IEEE, he has been researching the theory and practice of quantization and roundoff for the last 30 years. He is the author of about 135 scientific publications and has been involved in several industrial development projects.

Quantization Noise

Roundoff Error in Digital Computation, Signal Processing, Control, and Communications

Bernard Widrow and
István Kollár

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521886710

First published in print format 2008

ISBN-13 978-0-511-40611-9 eBook (Adobe Reader)

ISBN-13 978-0-521-88671-0 hardback

We dedicate this work to our fathers and our teachers. They influenced our lives and our thinking in a very positive way.

I would like to dedicate this book to Professors David Middleton and John G. Linvill, and to the memory of Professor William K. Linvill and my father, Moses Widrow.

Bernard Widrow

I would like to dedicate this book to Professors András Prékopa and Péter Osváth, and to the memory of Professor László Schnell and my father, Lajos Kollár.

István Kollár

Contents

Preface

For many years, rumors have been circulating in the realm of digital signal processing about quantization noise:

(a) the noise is additive and white and uncorrelated with the signal being quantized, and

(b) the noise is uniformly distributed between plus and minus half a quanta, giving it zero mean and a mean square of one-twelfth the square of a quanta.

Many successful systems incorporating uniform quantization have been built and placed into service worldwide whose designs are based on these rumors, thereby reinforcing their veracity. Yet simple reasoning leads one to conclude that:

(a) quantization noise is deterministically related to the signal being quantized and is certainly not independent of it,

(b) the probability density of the noise certainly depends on the probability density of the signal being quantized, and

(c) if the signal being quantized is correlated over time, the noise will certainly have some correlation over time.

In spite of the "simple reasoning," the rumors are true under most circumstances, or at least true to a very good approximation. When the rumors are true, wonderful things happen:

(a) digital signal processing systems are easy to design, and

(b) systems with quantization that are truly nonlinear behave like linear systems.

In order for the rumors to be true, it is necessary that the signal being quantized obeys a quantizing condition. There actually are several quantizing conditions, all pertaining to the probability density function (PDF) and the characteristic function (CF) of the signal being quantized. These conditions come from a "quantizing theorem" developed by B. Widrow in his MIT doctoral thesis (1956) and in subsequent work done in 1960.

Quantization works something like sampling, only the sampling applies in this case to probability densities rather than to signals. The quantizing theorem is related to the “sampling theorem,” which states that if one samples a signal at a rate at least twice as high as the highest frequency component of the signal, then the signal is recoverable from its samples. The sampling theorem in its various forms traces back to Cauchy, Lagrange, and Borel, with significant contributions over the years coming from E. T. Whittaker, J. M. Whittaker, Nyquist, Shannon, and Linvill.

Although uniform quantization is a nonlinear process, the “flow of probability” through the quantizer is linear. By working with the probability densities of the signals rather than with the signals themselves, one is able to use linear sampling theory to analyze quantization, a highly nonlinear process.

This book focuses on uniform quantization. Treatment of quantization noise, recovery of statistics from quantized data, analysis of quantization embedded in feedback systems, the use of “dither” signals and analysis of dither as “anti-alias filtering” for probability densities are some of the subjects discussed herein. This book also focuses on floating-point quantization which is described and analyzed in detail.

As a textbook, this book could be used as part of a mid-level course in digital signal processing, digital control, and numerical analysis. The mathematics involved is the same as that used in digital signal processing and control. Knowledge of sampling theory and Fourier transforms as well as elementary knowledge of statistics and random signals would be very helpful. Homework problems help instructors and students to use the book as a textbook.

Additional information is available from the following website:

`http://www.mit.bme.hu/books/quantization/`

where one can find data sets, some simulation software, generator programs for selected figures, etc. For instructors, the solutions of selected problems are also available for download in the form of a solutions manual, through the web pages above. It is desirable, however, that instructors also formulate specific problems based on their own experiences.

We hope that this book will be useful to statisticians, physicists, and engineers working in digital signal processing and control. We also hope that we have rescued from near oblivion some ideas about quantization that are far more useful in today’s digital world than they were when developed between 1955–60, when the number of computers that existed was very small. May the rumors circulate, with proper caution.

Acknowledgments

A large part of this book was written while István Kollár was a Fulbright scholar visiting with Bernard Widrow at Stanford University. His stay and work was supported by the Fulbright Commission, by Stanford University, by the US-Hungarian Joint Research Fund, and by the Budapest University of Technology and Economics. We gratefully acknowledge all their support.

The authors are very much indebted to many people who helped the creation of this book. Ideas described were discussed in different details with Tadeusz Dobrowiecki, János Sztipánovits, Ming-Chang Liu, Nelson Blachman, Michael Godfrey, László Györfi, Johan Schoukens, Rik Pintelon, Yves Rolain, and Tom Bryan. The ideas for some real-life problems came from András Vetier, László Kollár and Bernd Girod. Many exercises were taken from (Kollár, 1989). Valuable discussions were continued with the members of TC10 of IEEE's Instrumentation and Measurement Society, furthermore with the members of EUPAS (European Project for ADC-based devices Standardisation). Students of the reading classes EE390/391 of school years 2005/2006 and 2006/2007 at Stanford (Ekine Akuiyibo, Paul Gregory Baumstarck, Sudeepto Chakraborty, Xiaowei Ding, Altamash Janjua, Abhishek Prasad Kamath, Koushik Krishnan, Chien-An Lai, Sang-Min Lee, Sufeng Li, Evan Stephen Millar, Fernando Gomez Pancorbo, Robert Prakash, Paul Daniel Reynolds, Adam Rowell, Michael Shimasaki, Oscar Trejo-Huerta, Timothy Jwoyen Tsai, Gabriel Velarde, Cecylia Wati, Rohit Surendra Watve) pointed out numerous places to correct or improve.

The book could not have come to life without the continuous encouragement and help of Professor George Springer of Stanford University, and of Professor Gábor Péceli of the Budapest University of Technology and Economics.

A large fraction of the figures were plotted by Ming-Chang Liu, János Márkus, Attila Sárhegyi, Miklós Wágner, György Kálmán, and Gergely Turbucz. Various parts of the manuscript were typed by Mieko Parker, Joice DeBolt, and Patricia Halloran-Krokel. The LaTeX style used for typesetting was created by Gregory Plett. Very useful advice was given by Ferenc Wettl, Péter Szabó, László Balogh, and Zsuzsa Megyeri when making the final form of the book pages.

Last but not least, we would like to thank our families for their not ceasing support and their patience while enduring the endless sessions we had together on each chapter.

Glossary of Symbols

Throughout this book, a few formulas are repeated for easier reference during reading. In such cases, the repeated earlier equation number is typeset in italics, like in *(4.11)*.

a_k, b_k	Fourier coefficients
A	signal amplitude
A_{pp}	signal peak-to-peak amplitude
$\mathbf{A}^{\mathrm{T}}$	transpose of $\mathbf{A}$
$\mathbf{A}^*$	complex conjugate transpose of $\mathbf{A}$
$\overline{\mathbf{A}}$	complex conjugate of $\mathbf{A}$
B	bandwidth, or the number of bits in a fixed-point number (including the sign bit)
$\mathrm{cov}\{x, y\}$	covariance, page 42
$C(\tau)$	covariance function
d	dither, page 485
$\frac{\mathrm{d}x}{\mathrm{d}t}$	derivative
$\exp(\cdot)$	exponential function, also $e^{(\cdot)}$
$E(f)$	energy density spectrum
$\mathrm{E}\{x\}$	expected value (mean value)
f	frequency
f_{s}	sampling frequency, sampling rate
f_0	center frequency of a bandpass filter
f_1	fundamental frequency, or first harmonic
$f_x(x)$	probability density function (PDF), page 31
$F_x(x)$	probability distribution function, $F_x(x_0) = \mathrm{P}(x < x_0)$
$\Phi_x(u)$	characteristic function (CF): $\Phi_x(u) = \int_{-\infty}^{\infty} f_x(x) e^{jux}\,\mathrm{d}x = \mathrm{E}\{e^{jux}\}$ Eq. (2.17), page 27
$\mathcal{F}\{\cdot\}$	Fourier transform: $\mathcal{F}\{x(t)\} = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft}\,\mathrm{d}t$ for the PDF–CF pair, the Fourier transform is defined as $\int_{-\infty}^{\infty} f(x) e^{jux}\,\mathrm{d}x$

$\mathcal{F}^{-1}\{\cdot\}$	inverse Fourier transform: $\mathcal{F}^{-1}\{X(f)\} = \int_{-\infty}^{\infty} X(f) e^{j2\pi f t}\,\mathrm{d}f$ for the PDF–CF pair, the inverse Fourier transform is $\frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi(u)e^{-jux}\,\mathrm{d}u$
$h(t)$	impulse response
$H(f)$	transfer function
$\mathrm{Im}\{\cdot\}$	imaginary part
j	$\sqrt{-1}$
k	running index in time domain series
$\lg(\cdot)$	base-10 logarithm
$\ln(\cdot)$	natural logarithm (base e)
M_r	rth moment difference with PQN: $\mathrm{E}\{(x')^r\} - \mathrm{E}\{x^r\}$ Eq. (4.27), page 81
$\widetilde{M}_r$	rth centralized moment difference with PQN: $\mathrm{E}\{(\tilde{x}')^r\} - \mathrm{E}\{\tilde{x}^r\}$
n	pseudo quantization noise (PQN), page 69
n	frequency index (or: summation index in certain sums)
N	number of samples
N_r	small (usually negligible) terms in the rth moment: $\mathrm{E}\{(x')^r\} = \mathrm{E}\{x^r\} + M_r + N_r$, Eq. (B.1) of Appendix B, page 597
$\widetilde{N}_r$	small (usually negligible) terms in the rth centralized moment: $\mathrm{E}\{(\tilde{x}')^r\} = \mathrm{E}\{\tilde{x}^r\} + \widetilde{M}_r + \widetilde{N}_r$
$N(\mu,\, \sigma)$	normal distribution, page 49
$\mathcal{O}(x)$	decrease as quickly as x for $x \to 0$
p	precision in floating-point
p_i	probability
$\mathrm{P}\{\cdot\}$	probability of an event
q	quantum size in quantization, page 25
q_d	quantum size of a digital dither, page 686
q_{h}	step size of the hidden quantizer, page 357
Q	quality factor or weighting coefficient
$R(\tau)$	correlation function, Eq. (3.40), page 42
$R_{xy}(\tau)$	crosscorrelation function, $R_{xy}(\tau) = \mathrm{E}\{x(t)y(t+\tau)\}$ Eq. (3.41), page 42
R_r	residual error of Sheppard's rth correction Eq. (B.7) of Appendix B, page 602
$\widetilde{R}_r$	residual error of the rth Kind correction
$\mathrm{Re}\{\cdot\}$	real part
$\mathrm{rect}(z)$	rectangular pulse function, 1 if $\lvert z\rvert \le 0.5$, zero elsewhere
$\mathrm{rectw}(z)$	rectangular wave, 1 if $-0.25 \le z < 0.25$; -1 if $0.25 \le z < 0.75$; repeated with period 1
s	Laplace variable, or empirical standard deviation
s^*	corrected empirical standard deviation
S_r	Sheppard's rth correction, Eq. (4.29), page 82

$\widetilde{S}_r$	rth Kind correction
$S(f)$	power spectral density
$S_c(f)$	covariance power spectral density
$\mathrm{sign}(x)$	sign function
$\mathrm{sinc}(x)$	$\sin(x)/x$
T	sampling interval
T_m	measurement time
T_p	period length
T_r	record length
$\mathrm{tr}(z)$	triangular pulse function, $1 - \lvert z\rvert$ if $\lvert z\rvert \le 1$, zero elsewhere
$\mathrm{trw}(z)$	triangular wave, $1 - 4\lvert z\rvert$ if $\lvert z\rvert \le 0.5$, repeated with period 1
u	standard normal random variable
$u(t)$	time function of voltage
U	effective value of voltage
U_p	peak value
U_pp	peak-to-peak value
$\mathrm{var}\{x\}$	variance, same as square of standard deviation: $\mathrm{var}\{x\} = \sigma_x^2$
$w(t)$	window function in the time domain
$W(f)$	window function in the frequency domain
x	random variable
x'	quantized variable
$x' - x$	quantization noise, ν
$\tilde{x}$	centralized random variable, $x - \mu_x$, Eq. (3.13), page 34
$x(t)$	input time function
$X(f)$	Fourier transform of $x(t)$
$X(f, T)$	finite Fourier transform of $x(t)$
z^{-1}	delay operator, $e^{-j2\pi fT}$
δ	angle error
Δf	frequency increment, f_s/N in DFT or FFT
ϵ	error
ϵ_c	width of confidence interval
ϵ_r	relative error
φ	phase angle
$\gamma(f)$	coherence function: $\gamma(f) = \frac{S_{xy}(f)}{\sqrt{S_{xx}(f)S_{yy}(f)}}$
μ	mean value (expected value)
ν	quantization error, $\nu = x' - x$
Ψ	quantization fineness, $\Psi = 2\pi/q$
ω	radian frequency, $2\pi f$
Ω	sampling radian frequency, page 17
ρ	correlation coefficient (normalized covariance, $\frac{\mathrm{cov}\{x,y\}}{\sigma_x\sigma_y}$) Eq. (3.39), page 42

$\rho(t)$	normalized covariance function
σ	standard deviation
Σ	covariance matrix
τ	lag variable (in correlation functions)
ξ	$\xi = d + \nu$, total quantization error (in nonsubtractive dithering) Eq. (19.16), page 491
$\in$	element of set, value within given interval
$\star$	convolution: $\int_{-\infty}^{\infty} f(z)g(x-z)\,\mathrm{d}z = \int_{-\infty}^{\infty} f(x-z)g(z)\,\mathrm{d}z$
$\stackrel{\triangle}{=}$	definition
$\dot{\Phi}$	first derivative, e. g. $\dot{\Phi}_x(l\Psi) = \left.\frac{\mathrm{d}\,\Phi(u)}{\mathrm{d}(u)}\right\vert_{u=l\Psi}$
$\ddot{\Phi}$	second derivative, e. g. $\ddot{\Phi}_x(l\Psi) = \left.\frac{\mathrm{d}^2\,\Phi(u)}{\mathrm{d}(u)^2}\right\vert_{u=l\Psi}$
x'	quantized version of variable x
$\tilde{x}$	centralized version of variable x: $\tilde{x} = x - \mu_x$, Eq. (3.13), page 34
$\hat{x}$	estimated value of random variable x
$\lfloor x \rfloor$	nearest integer smaller than or equal to x (floor(x))
$\breve{x}$	deviation from a given value or variable

Acronyms and Abbreviations

AC	alternating current
ACF	autocorrelation function
A/D	analog-to-digital
ADC	analog-to-digital converter
AF	audio frequency (20 Hz–20 kHz)
AFC	automatic frequency control
AGC	automatic gain control
ALU	arithmetic and logic unit
AM	amplitude modulation
BW	bandwidth
CDF	cumulative distribution function
CF	characteristic function
CRT	cathode ray oscilloscope
D/A	digital-to-analog
DAC	digital-to-analog converter
dBV	decibels relative to 1 V
dBm	decibels relative to 1 mW
DC	direct current
DFT	discrete Fourier transform
DIF	decimation in frequency (a form of FFT algorithm)
DNL	differential nonlinearity
DIT	decimation in time (a form of FFT algorithm)
DSP	digital signal processing or digital signal processor
DUT	device under test
DVM	digital voltmeter
FIR	finite impulse response
FFT	fast Fourier transform
FM	frequency modulation
FRF	frequency response function (nonparametric)
HP	highpass (sometimes: Hewlett-Packard)
HV	high voltage

IC	integrated circuit
IF	intermediate frequency
IIR	infinite impulse response
INL	integral nonlinearity
I/O	input/output
LMS	least mean squares
LP	lowpass (*not* long play in this book)
LS	least squares
LSB	least significant bit
LSI	large scale integration
MAC	multiply and accumulate operation (A=A+B$*$C)
MSB	most significant bit
MUX	multiplexer
NFPQNP	normalized floating-point quantization noise power
NSR	noise-to-signal ratio, $10\log_{10}(P_{\mathrm{noise}}/P_{\mathrm{signal}})$
PDF	probability density function
PF	power factor
PLL	phase-locked loop
PQN	pseudo quantization noise
PSD	power spectral density (function)
PWM	pulse-width modulation
QT n	Quantizing Theorem n
RAM	random access memory
RC	resistance-capacitance (circuit)
RF	radio frequency
RMS	root mean square
ROM	read only memory
SEC	stochastic-ergodic converter
S/H	sample-hold
SI	Système International (d'Unités): International System of Units
SNR	signal-to-noise ratio, $10 \cdot \log_{10}(P_{\mathrm{signal}}/P_{\mathrm{noise}})$
SOS	sum of squares
TF	transfer function (parametric)
U/D	up/down
ULP	unit in the last place

Part I

Background

Chapter 1

Introduction

1.1 DEFINITION OF THE QUANTIZER

Quantization or roundoff occurs whenever physical quantities are represented numerically. The time displayed by a digital watch, the temperature indicated by a digital thermometer, the distances given on a map etc. are all examples of analog values represented by discrete numbers.

The values of measurements may be designated by integers corresponding to their nearest numbers of units. Roundoff errors have values between plus and minus one half unit, and can be made small by choice of the basic unit. It is apparent, however, that the smaller the size of the unit, the larger will be the numbers required to represent the same physical quantities and the greater will be the difficulty and expense in storing and processing these numbers. Often, a balance has to be struck between accuracy and economy. In order to establish such a balance, it is necessary to have a means of evaluating quantitatively the distortion resulting from rough quantization. The analytical difficulty arises from the inherent nonlinearities of the quantization process.

For purposes of analysis, it has been found convenient to define the quantizer as a nonlinear operator having the input-output staircase relation shown in Fig. 1.1(a). The quantizer output x' is a single-valued function of the input x, and the quantizer has an "average gain" of unity. The basic unit of quantization is designated by q. An input lying somewhere within a quantization "box" of width q will yield an output corresponding to the center of that box (i.e., the input is rounded-off to the center of the box). This quantizer is known as a "uniform quantizer."

The output of the quantizer will differ from the input. We will refer to this difference as ν, the "quantization noise," because in most cases it can be considered as a noise term added to the quantizer input. As such,

$$\nu = x' - x \, . \tag{1.1}$$

The quantizer symbol of Fig. 1.1(b) is useful in representing a rounding-off process with inputs and outputs that are signals in real time. As a mathematical operator, a

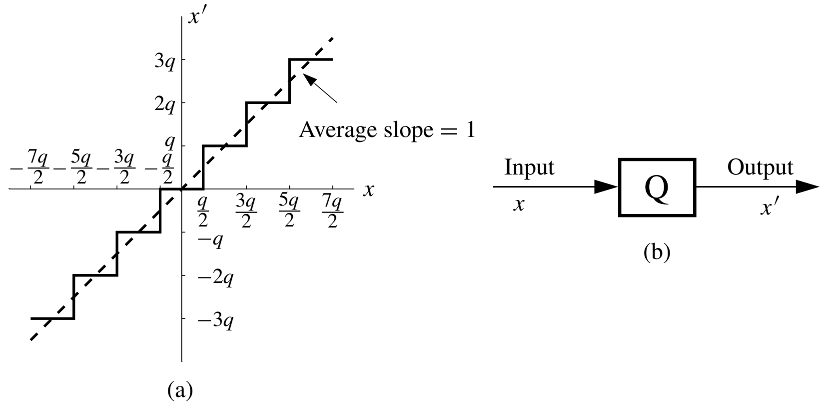

Figure 1.1 A basic quantizer (the so-called mid-tread quantizer, with a "dead zone" around zero): (a) input-output characteristic; (b) block-diagram symbol of the quantizer.

quantizer may be defined to process continuous signals and give a stepwise continuous output, or to process sampled signals and give a sampled output.

The attention of this work will be focused for the most part upon the basic quantizer of Fig. 1.1. The analysis that develops will be applicable to a variety of different kinds of quantizers which can be represented in terms of this basic quantizer and other simple linear and nonlinear operators. For example the quantizers shown in Fig. 1.2 and in Fig. 1.3 are derived from the basic quantizer by the addition of constants or dc levels to input and output, and by changing input and output scales, respectively. Notice that these input-output characteristics would approach the dotted lines whose slopes are the average gains if the quantization box sizes were made arbitrarily small.

Another kind of quantizer, one having hysteresis at each step, can be represented in terms of the basic quantizer with some positive feedback. The input-output characteristic is a staircase array of hysteresis loops. An example of this is shown in Fig. 1.4 for the quantization of both continuous and sampled signals. The average gain of this hysteresis quantizer is given by the feedback formula,

$$\begin{pmatrix} \text{average} \\ \text{gain} \end{pmatrix} = \frac{1}{1 - 1/4} = \frac{4}{3} \,. \tag{1.2}$$

Notice that a unit delay is included in the feedback loop of the sampled system. A unit delay (or more) must be incorporated within the feedback loop of any sampled system in order to avoid race conditions and to make feedback computation possible. The result of the delay in Fig. 1.4(c) is only to allow cycling of a hysteresis loop to take place from sample time to sample time.

Two- and three-level quantizers which are more commonly called saturating quantizers appear in nonlinear systems. They will be treated as ordinary quantizers

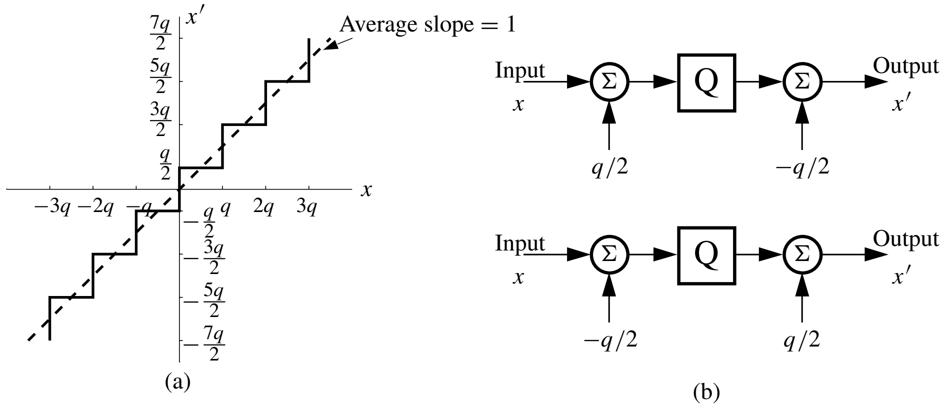

Figure 1.2 Effects of addition of constants: (a) a quantizer with comparison level at the origin, the so-called mid-riser quantizer (often used as a basic quantizer in control); (b) two equivalent representations of the characteristic in (a), using the basic quantizer defined in Fig. 1.1.

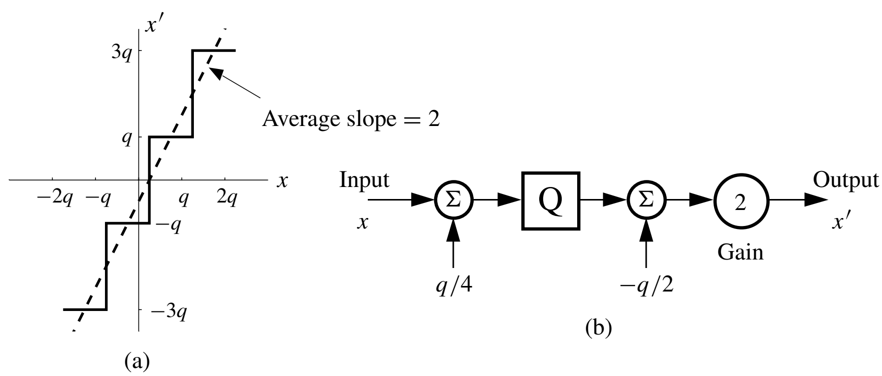

Figure 1.3 Effects of scale changes and addition of constants: (a) a quantizer with scale and dc level changes; (b) equivalent representation

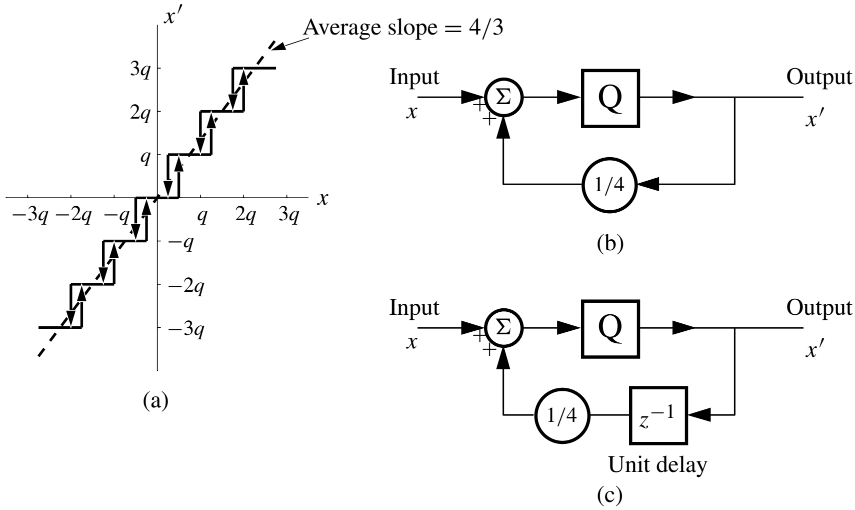

Figure 1.4 A quantizer with hysteresis: (a) input-output characteristic; (b) an equivalent representation for continuous signals; (c) an equivalent representation for sampled signals.

whose inputs are confined to two and three levels respectively. Fig. 1.5 shows their input-output characteristics and their block-diagram symbols. Fig. 1.6 shows examples of how saturating quantizers with hysteresis can be represented as saturating quantizers with positive feedback.

Every physical quantizer is noisy to a certain extent. By this is meant that the ability of the quantizer to resolve inputs which come very close to the box edges is limited. These box edges are actually smeared lines rather than infinitely sharp lines. If an input close to a box edge would be randomly rounded up or down, the quantizer could be represented as an ideal (infinitely sharp) basic quantizer with random noise added to its input (refer to Fig. 1.7).

Quantized systems result when quantizers are combined with dynamic elements. These systems may be open-looped or closed-looped, sampled or continuous, and linear or nonlinear (except for the quantizers). Quantized systems of many types will be discussed below.

Note that by changing the quantizer characteristics only slightly, like moving the position (the "offset") of the transfer characteristic along the dotted line, some properties of quantization will slightly change. The quantizer in Fig. 1.1(a), the so-called mid-tread quantizer, has a "dead zone" around zero. This quantizer is preferred by measurement engineers, since very small input values cause a stable zero output. We will execute derivations in this book usually assuming a mid-tread quantizer. On the other hand, the quantizer in Fig. 1.2(a) is the so-called mid-riser quantizer, with comparison level at zero. This is preferred by control engineers,

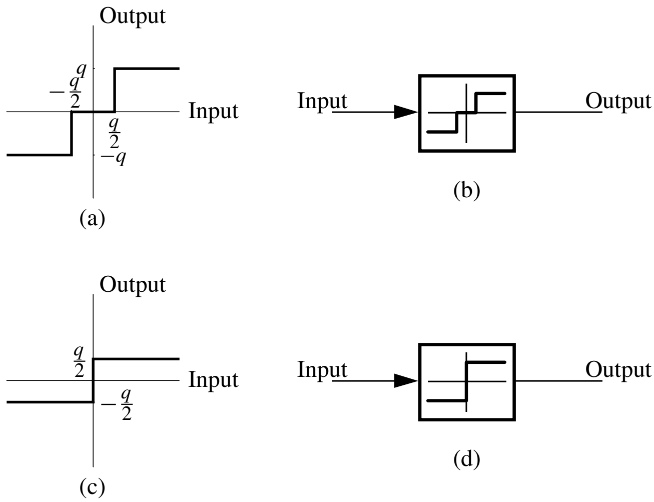

Figure 1.5 Saturating quantizers: (a) saturating quantizer with "dead zone"; (b) representation of (a); (c) the signum function, a saturating quantizer; (d) representation of (c).

because the output of this quantizer oscillates when its input oscillates around zero, allowing a feedback controller to force the measured quantity to zero on average, even with limited resolution.

There is another important aspect of quantization. Signal quantization occurs not only when analog quantities are transferred to a digital computer, but also occurs each time a calculated quantity is stored into the memory of a computer. This is called arithmetic rounding. It happens virtually at every step of calculations in the systems surrounding us, from mobile telephones to washing machines.

Arithmetic rounding is special in that the quantizer input is not an analog signal, but quantized data. For example, multiplication approximately doubles the number of bits (or that of the mantissa), and for storage we need to reduce the bit number back to that of the number representation. Thus, the number representation determines the possible quantizer outputs, and the rounding algorithm defines the quantizer transfer characteristic.

We cannot go into detailed discussion here of number representations. The interested reader is referred to (Oppenheim and Schafer, 1989). The main aspects are: the number representation can be fixed-point (uniform quantization) or floating-point (logarithmic or floating-point quantization, see later in Chapters 12 and 13). Coding of negative numbers can be sign-magnitude, two's complement or one's complement. Rounding can be executed to the nearest integer, towards zero, towards $\pm\infty$, upwards ("ceil") or downwards ("floor"). Moreover, finite bit length storage is often preceded by truncation (by simply dropping the excess bits), which leads to special transfer characteristics of the quantizer (see Exercise 1.6).

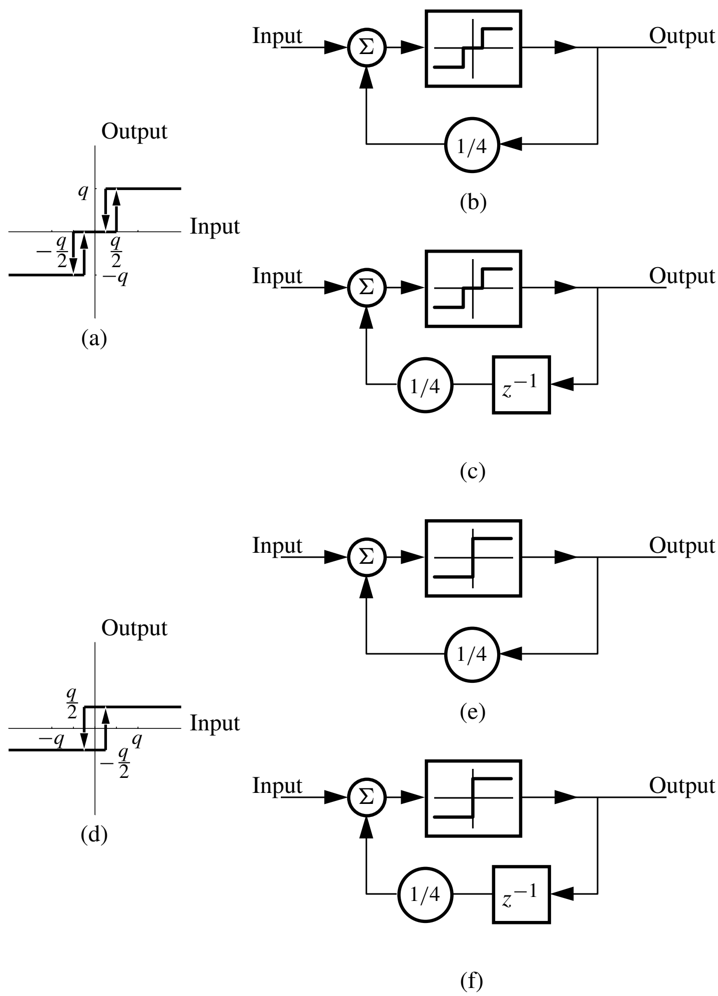

Figure 1.6 Saturating quantizers with hysteresis: (a) three-level saturating quantizer with hysteresis; (b) continuous-data representation of (a); (c) discrete-time representation of (a); (d) two-level saturating quantizer with hysteresis; (e) continuous-data representation of (d); (f) discrete-time representation of (d).

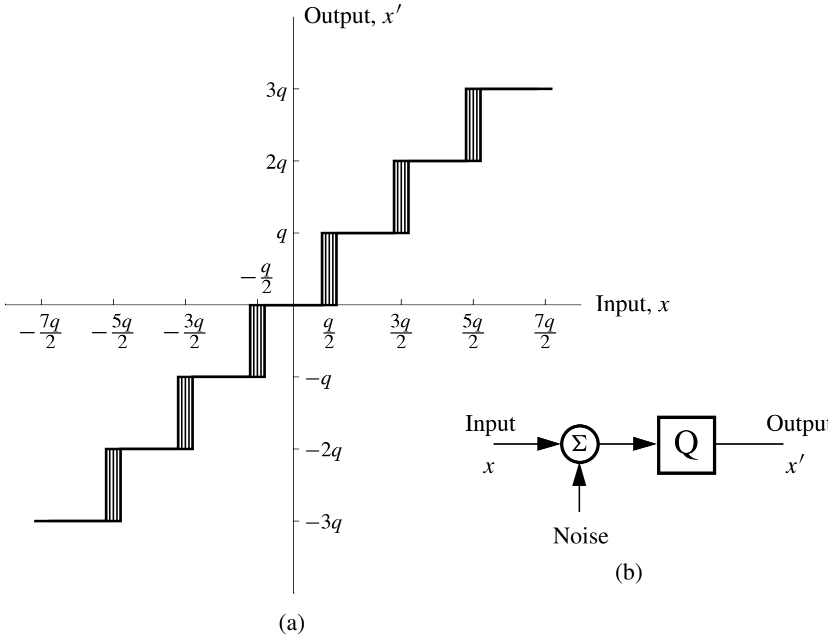

Figure 1.7 A noisy quantizer: (a) input-output characteristic; (b) representation of (a).

1.2 SAMPLING AND QUANTIZATION (ANALOG-TO-DIGITAL CONVERSION)

A numerical description of a continuous function of an independent variable may be made by plotting the function on graph paper as in Fig. 1.8. The function $x(t)$ can be approximately represented over the range $0 \leq t \leq 10$ by a series of numerical values, its quantized samples: 1, 3, 3, 2, 0, -1, -3, -3, -2, 0, 1.

The plot of Fig. 1.8 on a rectangular grid suggests that quantization in amplitude is somehow analogous to sampling in time. Quantization will in fact be shown to be a sampling process that acts not upon the function itself, however, but upon its probability density function.

Both sampling and quantization are effected when signals are converted from "analog-to-digital". Sampling and quantization are mathematically commutable operations. It makes no difference whether a signal is first sampled and then the samples are quantized, or if the signal is quantized and the stepwise continuous signal is then sampled. Both sampling and quantizing degrade the quality of a signal and may irreversibly diminish our knowledge of it.

A sampled quantized signal is discrete in both time and amplitude. Discrete systems behave very much like continuous systems in a macroscopic sense. They could be analyzed and designed as if they were conventional continuous systems by

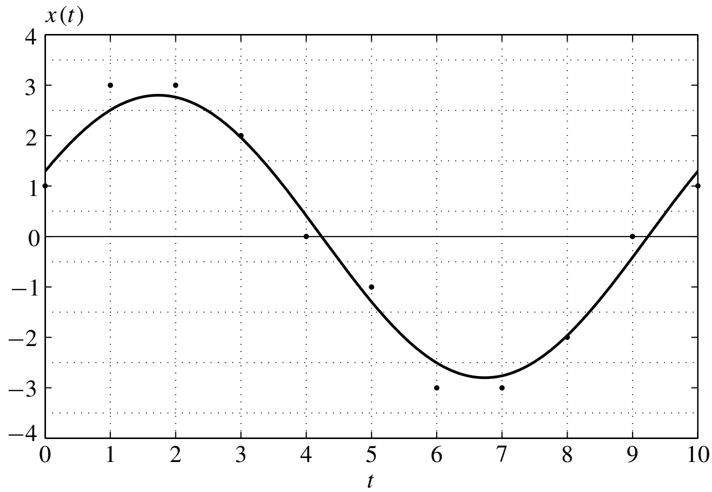

Figure 1.8 Sampling and quantization.

ignoring the effects of sampling. In order to take into account these effects, however, use must be made of sampling theory. Quantized systems, on the other hand, behave in a macroscopic sense very much like systems without quantization. They too could be analyzed and designed by ignoring the effects of quantization. These effects in turn could be reckoned with by applying the statistical theory of quantization. That is the subject of this book.

1.3 EXERCISES

1.1 Let the quantizer input x be the time function

$$\begin{array}{ll} x = 0, & t \le 0 \\ x = t, & 0 \le t \le 10 \\ x = 20 - t, & 10 \le t \le 20 \\ x = 0, & 20 \le t \end{array} . \tag{E1.1.1}$$

Let $q = 1$. Using Matlab, plot the quantizer output x' versus time

(**a**) for the quantizer of Fig. 1.1 (page 4),

(**b**) for the quantizer of Fig. 1.4(a), that is, of Fig. 1.4(c) (page 6). For the quantizer of Fig. 1.4(c), let the sampling period $T = 0.1$. Let the quantizer input x be samples of the continuous input of the above definition Eq. (E1.1.1).

1.2 Using Matlab, make a plot of the quantization error ν vs. input x,

(**a**) for the quantizer of Fig. 1.1 (page 4), $\nu = (x' - x)$,

(b) for the quantizer of Fig. 1.2 (page 5), $\nu = (x' - x)$,
(c) for the quantizer of Fig. 1.3(a) (page 5), $\nu = (x' - 2x)$,
(d) for the quantizer of Fig. 1.4(a) (page 6), $\nu = (x' - 4/3 \cdot x)$.

1.3 Finite resolution uniform quantization (with quantum step size q) can be simulated in Matlab, e.g. by any one of the following commands:

```
xq1=q*round(x/q); %Matlab's rounding
xq2=q*(x/q+pow2(53)-pow2(53)); %standard IEEE rounding
xq3=q*fix(x/q+sign(x)/2);
xq4=q*ceil(x/q-0.5);
xq5=q*fix( (ceil(x/q-0.5)+floor(x/q+0.5))/2 );
```

Do all of these expressions implement rounding to the closest integer? Are there differences among the results of these expressions? Why? What happens for the values of x such as -1.5, -0.5, 0.5, 1.5?

1.4 A crate of chicken bears on its tag the total weight rounded to the nearest pound. What is the maximum magnitude of the weight error in a truck load of 200 crates? (Remark: this bound is the so-called Bertram bound, see page 455.)

1.5 Two decimal numbers with number representation with two fractional digits (like in the number 74.52) are multiplied, and the result is stored after rounding to a similar form. Describe the equivalent quantizer characteristic. What is the corresponding quantum step size? What is the dynamic range[1] if two decimal digits are used for representing the integer part? How many quantum steps are included in the dynamic range?

Hint: look at the ratio of the largest and smallest representable positive values.

1.6 Number representations in digital systems, described by Oppenheim, Schafer and Buck (1998) and by other DSP texts, and by the world-wide web, correspond to certain quantizers. Draw the quantizer output vs. quantizer input for the following number representations:

(a) two's complement number representation,[2]
(b) one's complement number representation,[3]

[1] The dynamic range is a term used frequently in numerous fields to describe the ratio between the smallest and largest possible values of a changeable quantity, such as in sound and light.

[2] In two's complement representation, the leftmost bit of a signed binary numeral indicates the sign. If the leftmost bit is 0, the number is interpreted as a nonnegative binary number. If the most significant (leftmost) bit is 1, the bits contain a negative number in two's complement form. To obtain the absolute value of the negative number, all the bits are inverted, then 1 is added to the result.
A two's complement 8-bit binary numeral can represent any integer in the range -128 to $+127$. If the sign bit is 0, then the largest value that can be stored in the remaining seven bits is $2^7 - 1$, or 127. For example, $98 = 0110\,0010$, $-98 = 1001\,1110$.

[3] One's complement number representation is similar to two's complement, with the difference that in negative numbers, the bits of the absolute value are just inverted (no 1 is added). For example, $98 = 0110\,0010$, $-98 = 1001\,1101$.

(c) magnitude-and-sign number representation,

when the following algorithms are implemented:

i. rounding to the nearest integer,
ii. truncation,
iii. rounding towards zero,[4]
iv. rounding towards ∞ (upwards).[5]

Draw the quantizer output vs. the quantizer input for 4-bit numbers (sign bit + 3 bits), determine if the equivalent quantizer is uniform or not, and whether the quantizer is mid-tread or mid-riser (see Figs. 1.1, and 1.2).

[4]Rounding towards zero means that the number is rounded to the next possible rounded number in the direction of zero. For example, $1.9q$ is rounded to q, $1.2q$ is rounded to q, $-1.1q$ is rounded to $-q$, and $-1.8q$ is rounded to $-q$.

[5]Rounding towards ∞ means that the number is rounded to the next possible rounded number in the direction of $+\infty$. For example, $1.9q$ is rounded to $2q$, $1.1q$ is rounded to $2q$, $-1.1q$ is rounded to $-q$, and $-1.8q$ is rounded to $-q$.

Chapter 2

Sampling Theory

Discrete signals are sampled in time and quantized in amplitude. The granularity of such signals, caused by both sampling and quantization, can be analyzed by making use of sampling theory. This chapter reviews sampling theory and develops it in a conventional way for the analysis of sampling in time and for the description of sampled signals. Chapter 3 reviews basic statistical theory related to probability density, characteristic function, and moments. Chapter 4 will show how sampling theory and statistical ideas can be used to analyze quantization.

The origins of sampling theory and interpolation theory go back to the work of Cauchy, Borel, Lagrange, Laplace and Fourier, if not further. We do not have the space here to account for the whole history of sampling, so we will only highlight some major points. For historical details, refer to Higgins (1985), Jerri (1977), Marks (1991).

The sampling theorem, like many other fundamental theorems, was gradually developed by the giants of science, and it is not easy to determine the exact date of its appearance. Shannon (1949) remarks about the imprecise formulation of it that "this is a fact which is common knowledge in the communication art."

According to Higgins (1985), the first statement that is essentially equivalent to the sampling theorem is due to Borel (1897). The most often cited early paper is however the one of E. T. Whittaker (1915). He investigated the properties of the *cardinal function* which is the result of the interpolation formula used now in sampling theory, and showed that these are bandlimited. His results were further developed by his son, J. M. Whittaker (1929). J. M. Whittaker introduced the name *cardinal series* for the interpolation series.

The sampling theorem was first clearly formulated and brought to general knowledge in communication theory by Shannon (1949). It should be mentioned that Nyquist (1928) discussed topics very close to this 20 years before, and Kotel'nikov (1933) preceded Shannon by 15 years in formulating the sampling theorem, but the work of Kotel'nikov was not known outside the Soviet Union, while Shannon's work quickly became known throughout the world. Since Shannon's work, several different generalizations of the sampling theorem and the interpolation formula have been published. We need the sampling theorem for the analysis of quantization.

Sampling theory in a very useful form was developed by Linvill in his MIT doctoral thesis (Linvill, 1949). He derived expressions for the 2-sided Laplace transform and the Fourier transform of the output of a sampling device in terms of the Laplace and Fourier transforms of the input to the sampling device. He showed how sampling could be regarded as amplitude modulation of a periodic "impulse carrier" by the signal being sampled. Furthermore, he explained interpolation as filtering in the frequency domain, with an ideal lowpass filter providing sinc-function interpolation. An important paper by Linvill, based on his doctoral thesis, discussed the application of sampling theory to the analysis of discrete-time feedback control systems (Linvill, 1951). This work was further developed by Ragazzini and Zadeh (1952).

Widrow was exposed to the ideas of Linvill in 1952 when he took Linvill's MIT graduate course 6.54 "Pulse-Data Systems." In those days, graduate students at MIT were allowed to take up to 10% of their course program at Harvard. Taking advantage of this, Widrow took a course on statistical communication theory with Middleton at Harvard, first learning about probability densities and characteristic functions. Combining what was learned from Linvill and Middleton, Widrow did a doctoral thesis developing the statistical theory of quantization (Widrow, 1956a). The thesis was supervised by Linvill. This book is based on that doctoral thesis, and subsequent work.

The present chapter develops sampling theory from the point of view of Linvill. Although great work on the subject preceded Linvill, his ideas on the analysis of sampling in the frequency domain pervade today in the fields of digital signal processing and digital control and are second nature to scientists and engineers working in those fields. He analyzed sampling as a process of amplitude modulation.

It should be mentioned here that Linvill's approach makes use of an infinite series of Dirac delta functions. Such an impulse series is not a mathematical function, and conventional Laplace and Fourier theory does not apply to it. However, the whole derivation we are going to present can be rigorously justified by using distribution theory (Arsac, 1966; Bremermann, 1965; Lighthill, 1958; Zemanian, 1965).

2.1 LINVILL'S FREQUENCY DOMAIN DESCRIPTION OF SAMPLING

A simple derivation of sampling theory follows. Fig. 2.1(a) shows a continuous signal, the time function $x(t)$, being sampled. The samples are taken at uniform time intervals, each T seconds long. The samples may be mathematically represented by Dirac delta functions. The value (the area) of each delta function is made equal to the value of $x(t)$ at the given sampling instant multiplied by T. This scaling preserves the integral in the sense that the area of the samples of $x(t)$ approximately equals the area of $x(t)$.

In fact, when a signal is converted from analog to digital form, the result is a string of numbers representing the sampled values of $x(t)$. Physically, there are

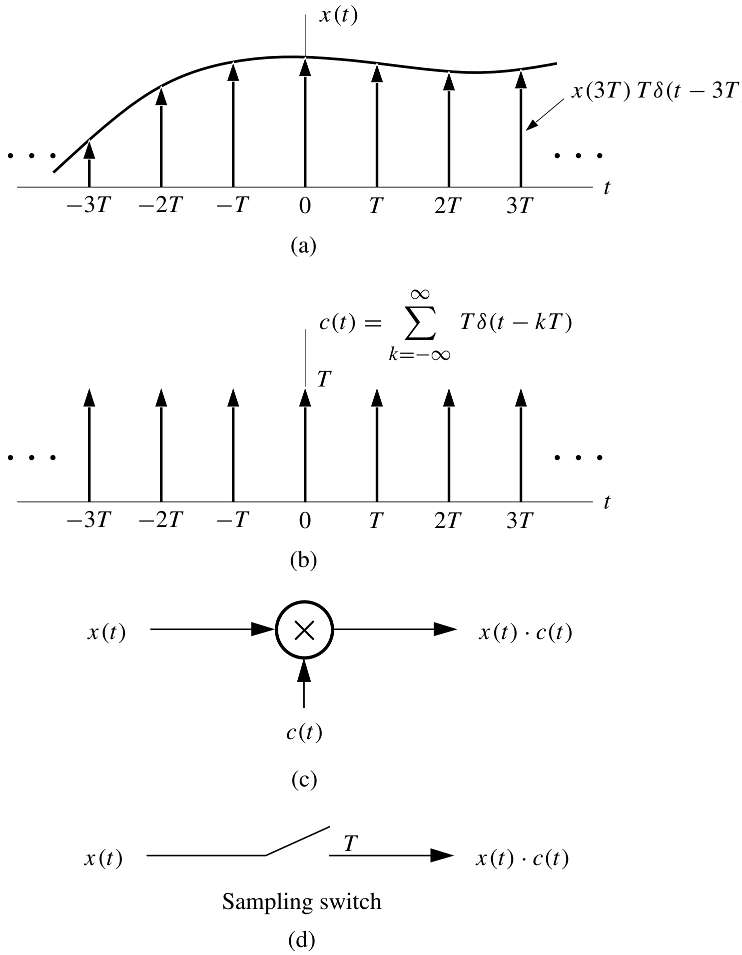

Figure 2.1 Sampling and amplitude modulation: (a) the signal and its samples; (b) the impulse carrier; (c) sampling represented as amplitude modulation; (d) sampling represented by the "sampling switch."

no impulses, just numbers that can be fed into a computer for numerical processing. Representation of the samples with Dirac delta functions is convenient here for purposes of analysis. Our goal is to be able to express the Fourier and Laplace transforms of the sampled signal in terms of the Fourier and Laplace transforms of the signal being sampled.

If the signal being sampled is $x(t)$, then the sampled signal will be a sum of delta functions or impulses given by

$$\sum_{k=-\infty}^{\infty} x(kT)\, T\delta(t-kT)\,. \tag{2.1}$$

This sampled time function could be obtained by multiplying $x(t)$ by a string of uniformly spaced impulses, each having area T.[1]

The signal $x(t)$ and its samples are shown in Fig. 2.1(a). A string of equal uniformly spaced impulses is shown in Fig. 2.1(b). It was called an "impulse carrier" by Linvill, and it can be represented as a sum of Dirac delta functions:

$$c(t) = \sum_{k=-\infty}^{\infty} T\delta(t-kT)\,. \tag{2.2}$$

Modulating (multiplying) this carrier with the signal $x(t)$ gives the samples of $x(t)$,

$$x(t)\cdot c(t) = \sum_{k=-\infty}^{\infty} x(kT)T\delta(t-kT)\,. \tag{2.3}$$

This multiplication or amplitude modulation is diagrammed in Fig. 2.1(c).

A representation of the sampling process commonly used in the literature is that of the "sampling switch," shown in Fig. 2.1(d). The input to the sampling switch is $x(t)$. Its output, the samples of $x(t)$, is given by $x(t)\cdot c(t)$.

[1] Since the only requirement for the Dirac delta series is that it represents the sample values, the scaling of the delta functions may be chosen arbitrarily. There are two popular ways for doing this. One possibility is to make the area of each delta function be equal to the sample value, which corresponds to a representation like $\sum x(kT)\delta(t-kT)$, with no extra scaling. The other possibility is to preserve the integral, by choosing $\sum x(kT)T\delta(t-kT)$. Since this is a more convenient choice for the representation of quantization, we decided to use it here.

The impulse carrier is periodic and representable as a Fourier series (Bracewell, 1986; Linvill, 1951; Papoulis, 1962).[2] The carrier can be expressed in terms of a complex Fourier series as

$$c(t) = \sum_{n=-\infty}^{\infty} a_n e^{jn\Omega t} . \tag{2.4}$$

The index n is the harmonic number. The nth Fourier coefficient is a_n. The sampling radian frequency is Ω. The sampling frequency in hertz is $\Omega/(2\pi)$. The sampling period is $T = 2\pi/\Omega$, so that $\Omega T = 2\pi$. The harmonic frequency is $n{\cdot}\Omega/(2\pi)$, a multiple of the sampling frequency.

The Fourier coefficient a_n is given by

$$a_n = \frac{1}{T} \int_{-T/2}^{T/2} c(t) e^{-jn\Omega t}\, \mathrm{d}t = \frac{1}{T} \int_{-T/2}^{T/2} T\delta(t) e^{-jn\Omega t}\, \mathrm{d}t = 1 . \tag{2.5}$$

Thus, the Fourier coefficient is real and has a unit value for all harmonics. Therefore, the impulse carrier can be represented by

$$c(t) = \sum_{n=-\infty}^{\infty} e^{jn\Omega t} . \tag{2.6}$$

Since the Fourier transform of each complex exponential is a Dirac delta function, the Fourier transform of a periodic series of Dirac delta functions is a sum of Dirac delta functions in the frequency domain:

$$\mathcal{F}\{c(t)\} = \mathcal{F}\left\{ \sum_{k=-\infty}^{\infty} T\delta(t-kT) \right\} = \sum_{n=-\infty}^{\infty} \delta(\omega - n\Omega) . \tag{2.7}$$

By the convolution theorem, the Fourier transform of a product of two signals is equal to the convolution of the two Fourier transforms. Therefore, we can directly write:

$$\mathcal{F}\big\{x(t)c(t)\big\} = X(j\omega) \star \sum_{n=-\infty}^{\infty} \delta(\omega - n\Omega) = \sum_{n=-\infty}^{\infty} X(j\omega - jn\Omega) . \tag{2.8}$$

This summation can be equivalently written as

$$\mathcal{F}\big\{x(t)c(t)\big\} = \sum_{n=-\infty}^{\infty} X(j\omega + jn\Omega) . \tag{2.9}$$

[2]This statement is not trivial. In strict mathematical sense, an infinite sum of periodically repeating Dirac delta functions is not a mathematical function, and thus the Fourier series cannot converge to it. Moreover, it cannot be "well" approximated by a finite Fourier series. However, in wider sense (distribution theory), development into a Fourier series is possible and the above statements are justified.

Figure 2.2 shows $x(t)$ being sampled. The input of the sampler is $x(t)$ in the time domain and $X(j\omega)$ in the Fourier domain. The output of the sampler is $x_s(t) =$

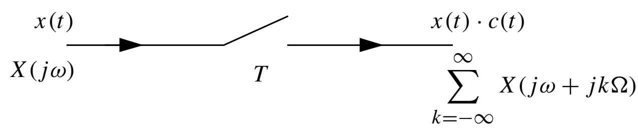

Figure 2.2 The signal $x(t)$ with transform $X(j\omega)$ being sampled.

$x(t)c(t)$ in the time domain, and in the Fourier domain, the output of the sampler is

$$X_s(j\omega) = \sum_{n=-\infty}^{\infty} X(j\omega + jn\Omega)\,. \tag{2.10}$$

It is easy to see how a signal transform is "mapped" through the sampling process. Fig. 2.3 illustrates this in the "frequency domain" for the Fourier transform, given by (2.9). Fig. 2.3(a) shows a continuous function of time $x(t)$ being sampled. The samples are $x(t)c(t)$. A symbolic representation of the Fourier transform of $x(t)$ is sketched in Fig. 2.3(b). This is $X(j\omega)$. Typically, $X(j\omega)$ is complex and has Hermitian symmetry about $j\omega = 0$. The Fourier transform of the samples is sketched in Fig. 2.3(c). This is the periodic function $\sum_{n=-\infty}^{\infty} X(j\omega + jn\Omega)$, and it is a sum of an infinite number of displaced replicas of $X(j\omega)$. These replicas, centered at $\omega = -\Omega$, $\omega = 0$, $\omega = \Omega$, and other multiples of Ω are shown in the figure. They correspond to $n = 1$, 0, and -1, etc., respectively.

2.2 THE SAMPLING THEOREM; RECOVERY OF THE TIME FUNCTION FROM ITS SAMPLES

If the replicas of $X(j\omega)$ in Fig. 2.3 contained in $\sum_{n=-\infty}^{\infty} X(j\omega + jn\Omega)$ do not overlap, it is possible to recover the original time function from its samples. Recovery of the original time function can be accomplished by lowpass filtering. Fig. 2.4(a) is a sketch of

$$\mathcal{F}\{x(t)c(t)\} = \sum_{n=-\infty}^{\infty} X(j\omega + jn\Omega)\,. \tag{2.9}$$

An ideal lowpass filter can separate the replica centered at $\omega = 0$ from all the others, and provide an output which is $X(j\omega)$ itself.

Define the ideal lowpass filter to have a gain of 1 in the passband $-\Omega/2 \leq \omega \leq \Omega/2$, and zero elsewhere. The frequency response of such a filter is sketched in

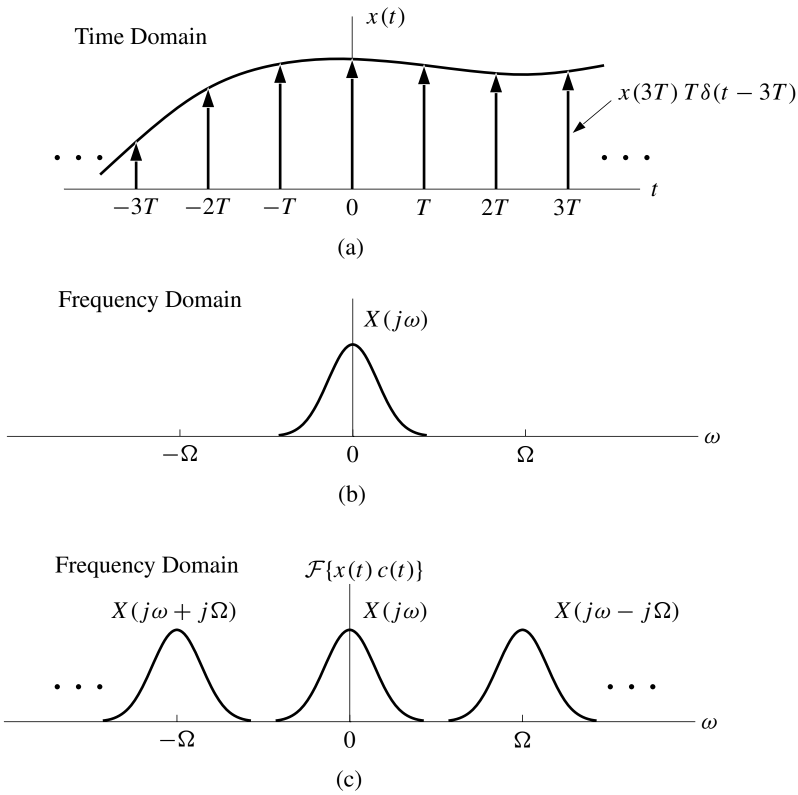

Figure 2.3 The Fourier transform of a time function, and the Fourier transform of its samples: (a) a time function being sampled; (b) Fourier transform of time function; (c) Fourier transform of samples of time function.

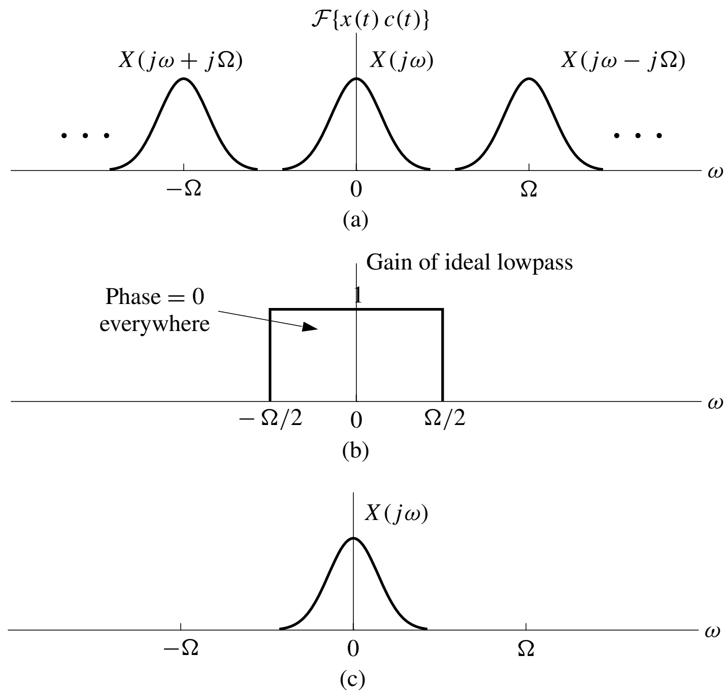

Figure 2.4 Recovery of original spectrum by ideal lowpass filtering of samples: (a) Fourier transform of samples of time function; (b) ideal lowpass filter; (c) Fourier transform of ideal lowpassed samples (the original time function).

Fig. 2.4(b). The Fourier transform of the output of the lowpass filter is sketched in Fig. 2.4(c). This is $X(j\omega)$, the Fourier transform of the original signal.

Based on the above considerations, we can state the sampling theorem.[3]

Sampling Theorem *If the sampling radian frequency Ω is high enough so that*

$$|X(j\omega)| = 0 \quad \textit{for} \quad |\omega| \geq \frac{\Omega}{2}, \tag{2.11}$$

then the sampling condition is met, and $x(t)$ is perfectly recoverable from its samples.

[3]The sampling theorem invariably holds in the case when none of the sampling instants is exactly at the origin (see Exercise 2.10). The phase term appearing in the repeated terms does not invalidate the argumentation.

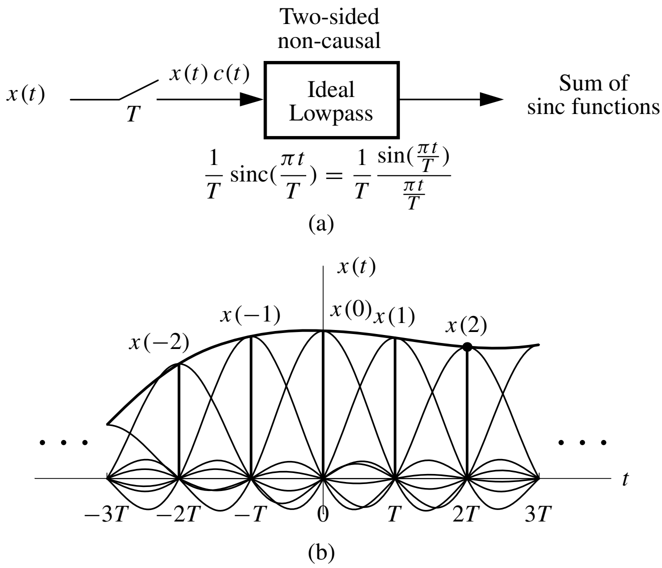

Figure 2.5 Ideal lowpass filtering (sinc function interpolation) for recovery of original signal from its samples: (a) ideal lowpass filtering of samples; (b) recovery of original time function by sinc function interpolation of its samples.

The replicas contained in $\sum_{n=-\infty}^{\infty} X(j\omega + jn\Omega)$ do not overlap when the sampling rate is at least twice as high as the highest frequency component contained in $x(t)$.

Recovering the function from its samples is illustrated in Fig. 2.5. In Fig. 2.5(a), $x(t)$ is sampled, and the resulting string of Dirac delta functions is applied to an ideal lowpass filter. At the filter output, the string of input delta functions is convolved with the impulse response of this filter. Since the transfer function $H(j\omega)$ of this filter is

$$H(j\omega) = \begin{cases} 1 & \text{if } |\omega| < \dfrac{\Omega}{2} \\ 0 & \text{if } |\omega| > \dfrac{\Omega}{2}, \end{cases} \tag{2.12}$$

the impulse response $h(t)$ is the inverse Fourier transform

$$h(t) = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} H(j\omega)e^{j\omega t}\,\mathrm{d}j\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(j\omega)e^{j\omega t}\,\mathrm{d}\omega$$

$$= \frac{1}{2\pi} \int_{-\Omega/2}^{\Omega/2} 1 \cdot e^{j\omega t}\, \mathrm{d}\omega = \left[\frac{e^{j\omega t}}{2\pi j t} \right]_{-\frac{\Omega}{2}}^{\frac{\Omega}{2}} = \frac{1}{T} \frac{\sin\left(\frac{\Omega}{2} t\right)}{\frac{\Omega}{2} t}$$

$$= \frac{1}{T} \operatorname{sinc}\left(\frac{\pi t}{T}\right). \tag{2.13}$$

Thus, the impulse response of the ideal lowpass filter is a sinc function, with a peak of amplitude $1/T$ at $t = 0$, and with zero crossings spaced every T seconds.

The output of the ideal lowpass filter, diagrammed in Fig. 2.5(a), will be the convolution of the sinc function (2.13) with the samples of $x(t)$. The idea is illustrated in Fig. 2.5(b):

$$x(t) = \sum_{k=-\infty}^{\infty} x(kT) \operatorname{sinc}\left(\pi \frac{t - kT}{T}\right). \tag{2.14}$$

The sum of the sinc functions will be exactly equal to $x(t)$ if the condition for the sampling theorem is met. This recovery of $x(t)$ from its samples is often called the interpolation formula, or sinc-function interpolation.

If the condition for the sampling theorem is not met, the interpolated function will be exactly equal to $x(t)$ only at the sampling instants. The spectral replicas will overlap in the frequency domain, and the ideal lowpass filter will not be able to extract $X(j\omega)$ without distortion.

From a historical perspective, it is interesting to note that the first sketches like those of Figs. 2.3–2.5 were drawn by Linvill. His work contributed to an intuitive understanding of sampling and interpolation. The mathematics had preceded him.

2.3 ANTI-ALIAS FILTERING

Fig. 2.6 illustrates the phenomenon of "aliasing" that occurs when sampling takes place and the condition for the sampling theorem is not satisfied. Aliasing is generally an undesirable effect and it can be prevented by "anti-alias filtering." Fig. 2.6(a) is a sketch of the Fourier transform of $x(t)$. Fig. 2.6(b) shows replicas of this transform, displaced by multiples of Ω over an infinite range of frequencies. It can be inferred from this diagram that components of $X(j\omega)$ whose frequencies are higher than $\Omega/2$ can in some cases be shifted in frequency to become components at frequencies lower than $\Omega/2$. This translation of high-frequency components into low-frequency components is called aliasing. Ideal lowpass filtering, illustrated in Fig. 2.6(c), cannot unscramble aliased frequency components. The result of such filtering is shown in Fig. 2.6(d). The output of the ideal lowpass filter is a distorted version of $x(t)$.

Aliasing can be prevented by making sure that the signal being sampled is band limited and that the sampling rate is high enough to satisfy the sampling condition. Fig. 2.7(a) illustrates the use of an anti-alias filter. The signal to be sampled $x(t)$

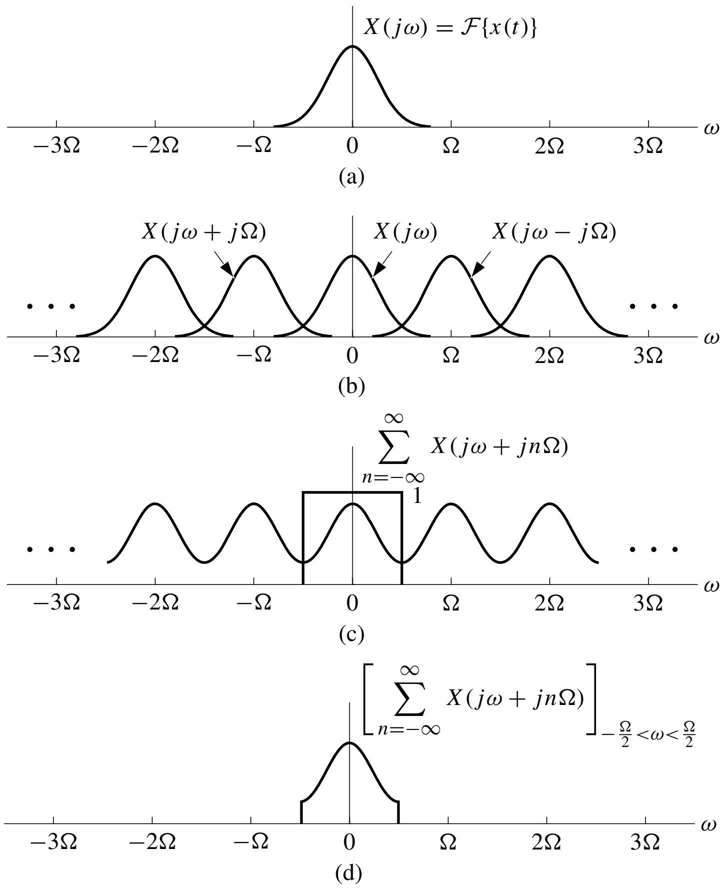

Figure 2.6 Difficulties in attempting to recover the original signal $x(t)$ from its aliased samples: (a) the spectrum of the original signal; (b) aliasing due to too-slow sampling; (c) attempt for recovery; (d) the result of recovery.

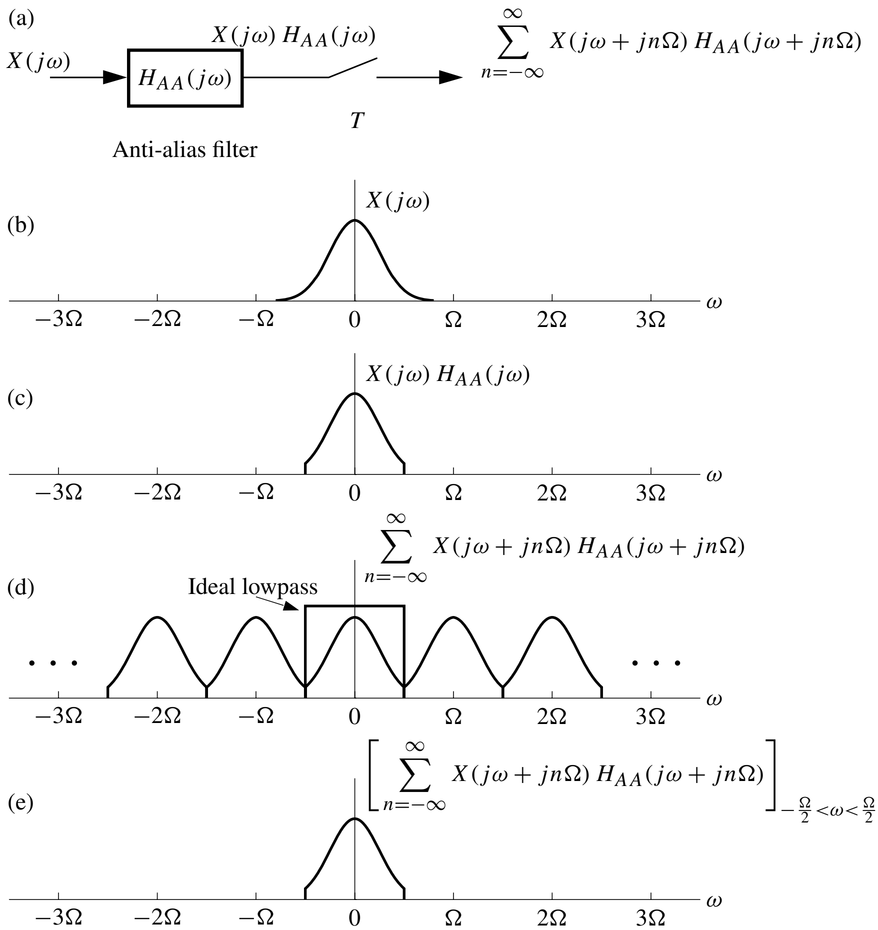

Figure 2.7 Use of anti-alias filtering before sampling: (a) block diagram; (b) original spectrum; (c) bandlimited spectrum; (d) repeated spectra; (e) recovered spectrum.

is applied as an input to a lowpass anti-alias filter whose output is then sampled. The Fourier transform of $x(t)$ is sketched in Fig. 2.7(b). In Fig. 2.7(c), the bandwidth of $X(j\omega)$ can be seen to have been reduced by the lowpass anti-alias filter. Subsequent sampling causes the spectral components of Fig. 2.7(c) to be repeated infinitely with spacing Ω. In this case, there is no overlap. Ideal lowpass filtering will not recover the original signal $x(t)$, but will yield $x(t)$ having gone through the anti-alias filter. The recovered signal will be a distorted version of $x(t)$ with some of its high-frequency components deleted, but with all of its frequency components occurring at the correct frequencies. In most circumstances, loss of high frequencies is preferable to having these frequencies appear later at lower frequencies.

Anti-alias filtering is widely practiced. A perfect anti-alias filter would be the ideal lowpass filter, having flat frequency response and zero phase shift in the range $-\Omega/2 < \omega < \Omega/2$, and zero response outside this range. Analog anti-alias filtering is always done whenever speech or music is recorded in digital form. Because practical analog filters do not allow sharp cutoff, analog filters followed by oversampling and digital lowpass filtering are generally used to do anti-alias filtering in order to make "clean" recordings.

2.4 A STATISTICAL DESCRIPTION OF QUANTIZATION, BASED ON SAMPLING THEORY

The analysis of quantization noise presented in this book is statistical in nature. In order to proceed with such an analysis, it is necessary to define the concepts of probability density, characteristic function, and moments. Although these are well-known ideas, a brief discussion seems appropriate at the outset.

Fig. 2.8 illustrates the probability density functions of the input and output of a quantizer. The quantizer is shown in Fig. 2.8(a). Its input is x, and its output is x'. The input–output relation is the stair-step function shown in Fig. 2.8(b). Fig. 2.8(c) shows the probability density function (PDF) of the quantizer input. This is represented by $f_x(x)$. The PDF of the quantizer output x' can be constructed from the PDF of the input x, as illustrated in the figure. The PDF of x' is represented by $f_{x'}(x')$.

When the input x occurs between $-q/2$ and $q/2$, the quantizer output x' has the value zero. The probability that this happens is

$$\int\limits_{-q/2}^{q/2} f_x(x)\,\mathrm{d}x\,. \tag{2.15}$$

This is therefore the probability that x' will have the exact value of zero. The probability density of x' will have a Dirac delta function at $x' = 0$ having an area equal to (2.15). In like manner, the PDF of x' can be constructed as a series of delta functions uniformly spaced with an interval of q. The area strips of $f_x(x)$ are compressed into delta functions to form $f_{x'}(x')$.

Accordingly,

$$\begin{aligned} f_{x'}(x') = \cdots + \delta(x'+q) \int\limits_{-\frac{3q}{2}}^{-\frac{q}{2}} f_x(x)\,\mathrm{d}x + \delta(x') \int\limits_{-\frac{q}{2}}^{\frac{q}{2}} f_x(x)\,\mathrm{d}x \\ + \delta(x'-q) \int\limits_{\frac{q}{2}}^{\frac{3q}{2}} f_x(x)\,\mathrm{d}x + \cdots \end{aligned} \tag{2.16}$$

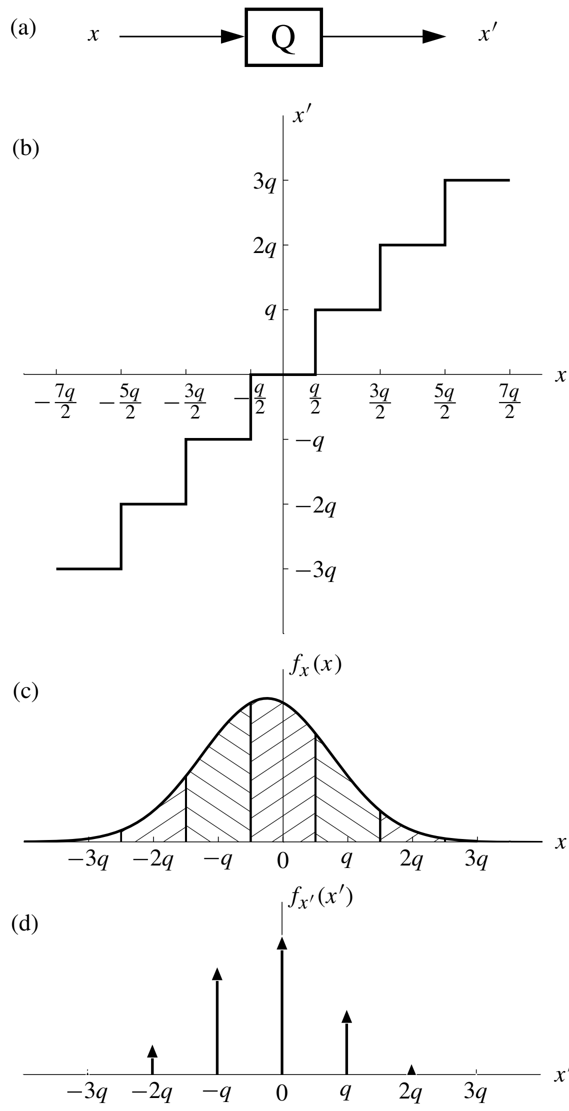

Figure 2.8 The probability density functions of the input and output of a quantizer: (a) the quantizer; (b) stair-step input–output characteristic; (c) PDF of x; (d) PDF of x'.

Forming $f_{x'}(x')$ from $f_x(x)$ looks like a sampling process. The continuous PDF of x at the quantizer input maps into the discrete PDF of x'. This is indeed a sampling process, but it is different from the one shown in Fig. 2.1. We need to "rotate our thinking by 90 degrees".

The sample values of $f_{x'}(x')$ do not relate to sample values of $f_x(x)$, but to the areas of the corresponding strips of $f_x(x)$. This kind of sampling is called "area sampling" (Widrow, 1956a; Widrow, 1956b), and will be described further in Chapter 4.

In this chapter, sampling of signals was studied by using the Fourier transforms of the input and output signals of the sampling switch. Linear methods were used, since sampling is a linear process. Although quantization is a nonlinear process with a stair-step input-output relation, the mapping of probability density through it is a linear process, i.e., area sampling is linear. Linear methods based on Fourier transformation of $f_x(x)$ and $f_{x'}(x')$ are used to analyze quantization.

The Fourier transform of the PDF is called its "characteristic function" (CF). The CF of input x is defined as[4]

$$\Phi_x(u) \stackrel{\triangle}{=} \int_{-\infty}^{\infty} f_x(x) e^{jux}\,\mathrm{d}x = \mathrm{E}\left\{e^{jux}\right\} . \tag{2.17}$$

Therefore, the modulation property is the same as we are accustomed to when using variable $\omega = 2\pi f$ in the frequency domain. The Fourier transform of the product of two functions is the convolution of their transforms:

$$\int_{-\infty}^{\infty} f_{x_1}(x) f_{x_2}(x) e^{jux}\,\mathrm{d}x = \frac{1}{2\pi}\left(\Phi_{x_1}(u) \star \Phi_{x_2}(u)\right) . \tag{2.18}$$

The CF of the quantizer output x' is obtained by Fourier transforming $f_{x'}(x')$, which is represented by (2.16). The result is

$$\Phi_{x'}(u) = \cdots + e^{juq} \int_{-\frac{3q}{2}}^{-\frac{q}{2}} f_x(x)\,\mathrm{d}x + \int_{-\frac{q}{2}}^{\frac{q}{2}} f_x(x)\,\mathrm{d}x + e^{-juq} \int_{\frac{q}{2}}^{\frac{3q}{2}} f_x(x)\,\mathrm{d}x + \cdots \tag{2.19}$$

We would like to find a circumspect relationship between (2.17) and (2.19) which would in some way be analogous to the relationship between the Fourier transform $X(j\omega)$ of signal $x(t)$ and the Fourier transform of its samples,

$$X_{\mathrm{s}}(j\omega) = \sum_{n=-\infty}^{\infty} X(j\omega + jn\Omega) . \tag{2.10}$$

[4]This is the generally adopted definition of the CF. In engineering fields, the Fourier transform is generally defined with a negative exponent. Apart from this difference, all properties of Fourier transform pairs hold for the PDF–CF pair, and a corresponding "sampling theorem" can also be developed.

We are looking for a quantizing theorem for signal quantization which would be analogous to the sampling theorem for signal sampling. To achieve our goal, we will need to take an aside and review some basic notions of statistics, probability density functions, cumulative distribution functions, characteristic functions, and moments. This is the subject of Chapter 3.

2.5 EXERCISES

2.1 We are looking through a picket fence at the shape of a house with a high roof. We see the total width of 46 feet through 24 vertical slots. The slope of the roof is 3/4, and the peak is half way between slots. The minimum height of the roof is 10 feet. What can we say about the precision of our knowledge of the top of the triangle-shaped roof, if

- (**a**) we can use sinc interpolation of the samples to determine its position? What will the error be? Calculate by computer.
- (**b**) we use the fact that the contour of the house and its roof consists of straight lines? What will the error be?

2.2 We would like to sample a sine wave of frequency $f_1 \approx 100\,\text{MHz}$ in such a way that an alias product appears at about $f = 100\,\text{kHz}$.

- (**a**) How shall we choose the sampling frequency?
- (**b**) Illustrate the spectrum of the sampled signal.

2.3 In a historical film we observe that the wheels of the coach seem to turn slowly backwards, then they stop and begin to turn slowly forwards, making one turn in $T = 3\,\text{s}$. What is the speed of the coach? What is the speed if the number of the spokes seems to be doubled? Basic data: The frames of the film are taken at a rate of $f = 24/\text{s}$; the diameter of the wheel of the coach is $d = 1$ m, the number of the spokes is $k = 12$.

2.4 The bell curve (the Gaussian density function) is defined as:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} . \tag{E2.4.1}$$

Imagine the Gaussian curve as if it were a time function.

- (**a**) Give the sampling distances prescribed by the sampling theorem (a) theoretically, (b) approximately (make a reasonable approximation).
- (**b**) Determine the Fourier transform of the sampled bell curve if the samples are taken at uniform distances d.
- (**c**) Estimate the maximum reconstruction error if $d = \sigma$. Calculate by computer.

2.5 The output of a very selective bandpass filter is to be sampled. The passband is somewhere between 1 kHz and 1.15 kHz, outside these frequency limits the attenuation is strong.

- (**a**) Calculate the value of the Nyquist frequency for the output signal of the filter.

(b) Give a rough lower bound for the sampling frequency, providing that the signal can be reconstructed from the samples.

(c) Choosing the sampling frequency equal to this lower bound, check if the signal can be actually reconstructed or not. Determine the minimum sampling frequency which allows signal recovery.

(d) Give a general rule to sample the output signal of a bandpass filter in (f_1, f_2): How can the minimum sampling frequency be determined? How can it be checked to see whether a given sampling frequency is appropriate or not?

2.6 Prove that the power $P = U_{\mathrm{eff}} I_{\mathrm{eff}} \cos\varphi$ (U_{eff} is the effective (RMS) value of the voltage, I_{eff} is the effective (RMS) value of the current, and φ is the phase shift between the voltage and the current) in a harmonic system (e. g. in the power mains) can be exactly determined from voltage and current samples by evaluating the expression

$$P = \frac{1}{N} \sum_{k=1}^{N} u_k i_k \tag{E2.6.1}$$

where u_k, i_k are periodic samples of the voltage and the current, respectively, if for the sampling the following equation is valid:

$$N T_{\mathrm{s}} = M T_{\mathrm{p}}\,, \tag{E2.6.2}$$

with

$$\begin{aligned} &N \quad \text{is the number of samples } (N > 0),\\ &T_{\mathrm{s}} \quad \text{is the sampling distance},\\ &M \quad \text{is a positive integer},\\ &2M/N \quad \text{is not an integer},\\ &T_{\mathrm{p}} = 1/f_1 \quad \text{is the period length of the sine wave}\,. \end{aligned} \tag{E2.6.3}$$

Notice that if $N < 2M$, the condition of the sampling theorem, $f_{\mathrm{s}} > 2f_1$, is not fulfilled for the voltage and current signals, and if $N < 4M$, the instantaneous power at frequency $f_{\mathrm{p}} = 2f_1$ is not sampled properly either. Considering these facts, discuss the above statement and its relation to the sampling theorem.

2.7 The spectrum of an almost symmetric square wave (fill-in factor ≈ 0.5) of frequency $f_1 \approx 100\,\mathrm{Hz}$, is measured using a Fourier analyzer. How is the sampling frequency to be chosen if the displaying is linear, and the vertical resolution of the display is approximately 1%?

2.8 Prove that if an arbitrary transient signal is sampled, an upper bound of the error of reconstruction is:

$$|\hat{x}(t) - x(t)| \le \frac{1}{\pi} \int_{-\infty}^{-\Omega_{\mathrm{s}}/2} |X(\omega)|\,\mathrm{d}\omega + \frac{1}{\pi} \int_{\Omega_{\mathrm{s}}/2}^{\infty} |X(\omega)|\,\mathrm{d}\omega \tag{E2.8.1}$$

where

$$\begin{aligned} X(\omega) &= \mathcal{F}\{x(t)\}\,, \\ \hat{x}(t) &= \sum_{k=-\infty}^{\infty} x(kT_\mathrm{s})\,\mathrm{sinc}(\Omega_\mathrm{s} t/2 - k\pi)\,, \\ \mathrm{sinc}(x) &= \frac{\sin x}{x}\,, \\ \Omega_\mathrm{s} &= \frac{2\pi}{T_\mathrm{s}}\,. \end{aligned} \tag{E2.8.2}$$

2.9 Give an estimate for the number of samples necessary for the reconstruction of the functions

$$x_1(t) = e^{-\frac{|t|}{T}}, \qquad \text{and} \qquad x_2(t) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{(t-\mu)^2}{2\sigma^2}}, \tag{E2.9.1}$$

respectively, if the error, relative to the maximum value, should not be larger than 0.01 . Hint: use the bound given in Exercise 2.8.

2.10 The sampling theorem was stated for the case when one of the sampling instants coincides with the origin.

(a) How will the Fourier transform of the sampled signal, given in Eq. (2.10), change when the sampling is delayed by time T_d?

(b) Does this influence the validity condition of the sampling theorem?

(c) How does the Fourier transform of the sampled signal change if the sampling instants remain centralized, but the signal is delayed by time $T_{\mathrm{d}2}$?

2.11 The proof of the interpolation formula (2.14) is based on the selection of the central replica of the original spectrum by multiplication of the Fourier transform of the sampled signal by a rectangular window function. However, the rectangular shape of the window is not a must. If the signal is more severely bandlimited (e.g. $X(j\omega) = 0$ for $|\omega| \geq \Omega_\mathrm{b}/2$, where $\Omega_\mathrm{b} < \Omega$, it is sufficient that the window is constant for frequencies where the Fourier transform of the input signal is nonzero.

(a) Generalize the interpolation formula for this case.

(b) This suggests a possibility to accelerate the convergence of the formula. How would this work?

Chapter 3

Probability Density Functions, Characteristic Functions, and Moments

The purpose of this chapter is to provide an introduction to the basics of statistical analysis, to discuss the ideas of probability density function (PDF), characteristic function (CF), and moments. Our goal is to show how the characteristic function can be used to obtain the PDF and moments of functions of statistically related variables. This subject is useful for the study of quantization noise.

3.1 PROBABILITY DENSITY FUNCTION

Figure 3.1(a) shows an ensemble of random time functions, sampled at time instant $t = t_1$ as indicated by the vertical dashed line. Each of the samples is quantized in amplitude. A "histogram" is shown in Fig. 3.1(b). This is a "bar graph" indicating the relative frequency of the samples falling within the given quantum box. Each bar can be constructed to have an area equal to the probability of the signal falling within the corresponding quantum box at time $t = t_1$. The sum of the areas must total to 1. The ensemble should have an arbitrarily large number of member functions. As such, the probability will be equal to the ratio of the number of "hits" in the given quantum box divided by the number of samples. If the quantum box size is made smaller and smaller, in the limit the histogram becomes $f_x(x)$, the probability density function (PDF) of x, sketched in Fig. 3.1(c). The area under the PDF curve is 1:

$$\int_{-\infty}^{\infty} f_x(x)\,\mathrm{d}x = 1\,. \tag{3.1}$$

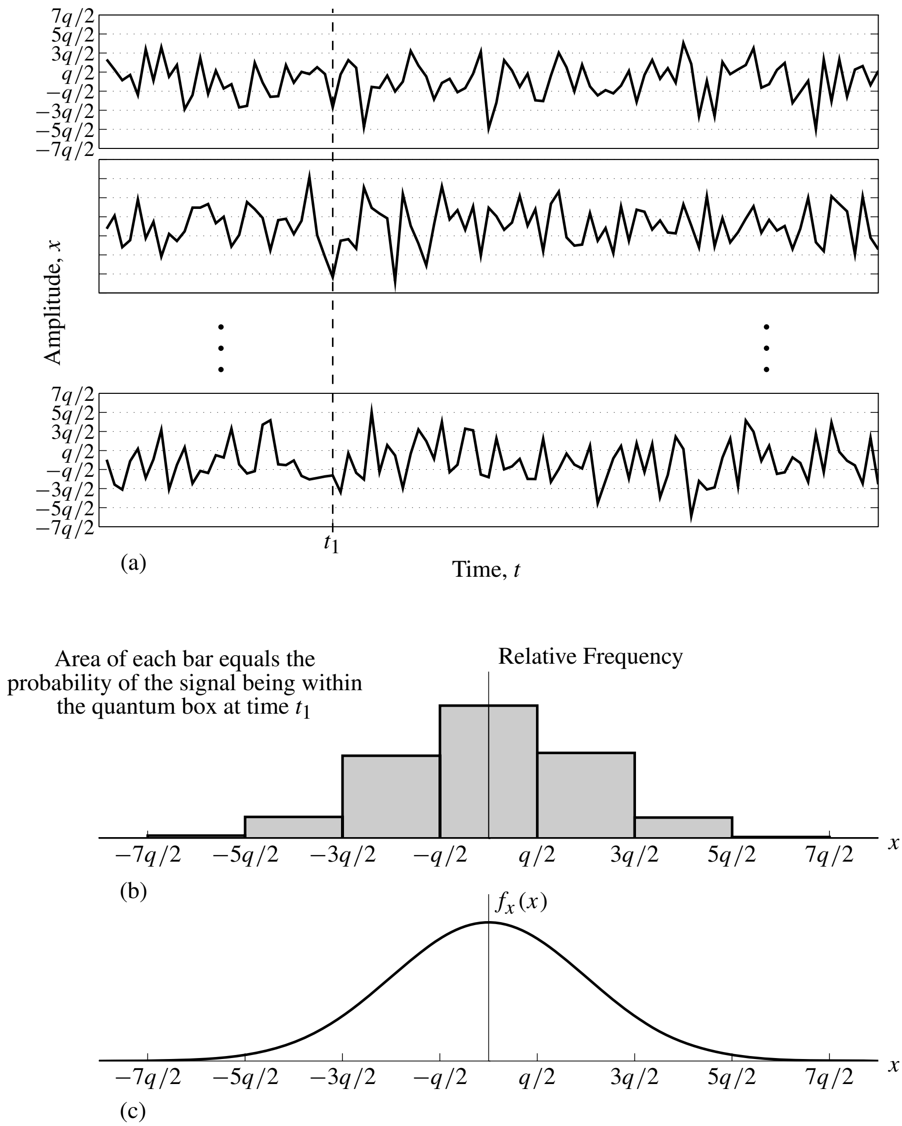

Figure 3.1 Derivation of a histogram: (a) an ensemble of random time functions; (b) a histogram of x; (c) the PDF of x.

The integral

$$\int_{x_0-\frac{\Delta x}{2}}^{x_0+\frac{\Delta x}{2}} f_x(x)\,\mathrm{d}x \tag{3.2}$$

gives the probability that the amplitude of the random variable falls in the interval $(x_0 - \frac{\Delta x}{2}, x_0 + \frac{\Delta x}{2})$.

3.2 CHARACTERISTIC FUNCTION AND MOMENTS

The PDF of x has the characteristic function (CF) given by Eq. (2.17)

$$\Phi_x(u) = \int_{-\infty}^{\infty} f_x(x) e^{jux}\,\mathrm{d}x\,. \qquad \textit{(2.17)}$$

The value of the characteristic function at $u = 0$, the origin in the "characteristic function domain," is

$$\Phi_x(0) = \int_{-\infty}^{\infty} f_x(x)\,\mathrm{d}x = 1\,. \tag{3.3}$$

The value of a characteristic function at its origin must always be unity.

Since $\Phi_x(u)$ as given in (2.17) is the Fourier transform of $f_x(x)$, and since $f_x(x)$ is real, the characteristic function is conjugate symmetric:

$$\Phi_x(-u) = \overline{\Phi_x(u)}\,. \tag{3.4}$$

The overline denotes the complex conjugate.

The values of the derivatives of the characteristic function at the origin are related to moments of x. Let us assume in the following paragraphs that the moments we are investigating exist. For the CF of Eq. (2.17), differentiation yields

$$\frac{\mathrm{d}\,\Phi_x(u)}{\mathrm{d}u} = \dot{\Phi}_x(u) = \int_{-\infty}^{\infty} jx f_x(x) e^{jux}\,\mathrm{d}x\,. \tag{3.5}$$

Note that derivatives of various orders are indicated in Newtonian fashion by dots over the relevant variables.

Evaluation of the derivative (3.5) at $u = 0$ gives

$$\left.\frac{\mathrm{d}\,\Phi_x(u)}{\mathrm{d}u}\right|_{u=0} = \dot{\Phi}_x(0) = \int\limits_{-\infty}^{\infty} jx f_x(x)\,\mathrm{d}x = j\mathrm{E}\{x\}\,. \tag{3.6}$$

Taking the kth derivative of (2.17) gives

$$\frac{\mathrm{d}^k\,\Phi_x(u)}{\mathrm{d}u^k} = \int\limits_{-\infty}^{\infty} (jx)^k f_x(x) e^{jux}\,\mathrm{d}x\,. \tag{3.7}$$

Evaluating this at $u = 0$,

$$\left.\frac{\mathrm{d}^k\,\Phi_x(u)}{\mathrm{d}u^k}\right|_{u=0} = \int\limits_{-\infty}^{\infty} (jx)^k f_x(x)\,\mathrm{d}x = j^k\mathrm{E}\{x^k\}\,. \tag{3.8}$$

Accordingly, the kth moment of x is

$$\mathrm{E}\{x^k\} = \frac{1}{j^k}\left.\frac{\mathrm{d}^k\,\Phi_x(u)}{\mathrm{d}u^k}\right|_{u=0}\,. \tag{3.9}$$

It is sometimes useful to be able to find the moments of $g(x)$, a function of x. The first moment is

$$\mathrm{E}\{g(x)\} = \int\limits_{-\infty}^{\infty} g(x) f_x(x)\,\mathrm{d}x\,. \tag{3.10}$$

An important function of x is e^{jux}. We can obtain the first moment of this function as

$$\mathrm{E}\{e^{jux}\} = \int\limits_{-\infty}^{\infty} f_x(x) e^{jux}\,\mathrm{d}x\,. \tag{3.11}$$

This moment can be recognized to be the CF of x. Accordingly,

$$\mathrm{E}\{e^{jux}\} = \int\limits_{-\infty}^{\infty} f_x(x) e^{jux}\,\mathrm{d}x = \Phi_x(u)\,. \tag{3.12}$$

We will represent a zero-mean version of x as $\tilde{x}$, the difference between x and its mean value $\mu = \mathrm{E}\{x\}$:

$$\tilde{x} \stackrel{\triangle}{=} x - \mu\,. \tag{3.13}$$

The characteristic function of $\tilde{x}$ can be obtained from

$$\Phi_x(u) = \mathrm{E}\{e^{jux}\} = \mathrm{E}\{e^{ju(\tilde{x}+\mu)}\} = e^{ju\mu}\, \Phi_{\tilde{x}}(u)\,, \tag{3.14}$$

$$\Phi_{\tilde{x}}(u) = e^{-ju\mu}\, \Phi_x(u)\,. \tag{3.15}$$

By using this expression, the kth derivative of the characteristic function is

$$\frac{\mathrm{d}^k\, \Phi_x(u)}{\mathrm{d}u^k} = \frac{\mathrm{d}^k}{\mathrm{d}u^k}\left(e^{ju\mu}\, \Phi_{\tilde{x}}(u)\right)\,. \tag{3.16}$$

For $k = 1$ this gives the following:

$$\frac{\mathrm{d}\, \Phi_x(u)}{\mathrm{d}u} = \dot{\Phi}_x(u) = e^{ju\mu}\left(j\mu\, \Phi_{\tilde{x}}(u) + \dot{\Phi}_{\tilde{x}}(u)\right)\,. \tag{3.17}$$

From the above expression for the kth derivative, the kth moment of x is

$$\begin{aligned} \mathrm{E}\{x^k\} &= \frac{1}{j^k}\frac{\mathrm{d}^k\, \Phi_x(u)}{\mathrm{d}u^k}\bigg|_{u=0} \\ &= \frac{1}{j^k}\left[\frac{\mathrm{d}^k}{\mathrm{d}u^k} e^{ju\mu}\, \Phi_{\tilde{x}}(u)\right]_{u=0}\,. \end{aligned} \tag{3.18}$$

We will use this expression subsequently. For $k = 1$, this expression reduces to

$$\mathrm{E}\{x\} = \frac{\dot{\Phi}_x(0)}{j} = \left(\mu\, \Phi_{\tilde{x}}(0) + \frac{\dot{\Phi}_{\tilde{x}}(0)}{j}\right) = \mu\,. \tag{3.19}$$

3.3 JOINT PROBABILITY DENSITY FUNCTIONS

An ensemble of random time functions which are sampled at times $t = t_1$ and $t = t_2$ is shown in Fig.3.2(a). The joint probability density function of the samples x_1 and x_2, taken at times t_1 and t_2, is represented by $f_{x_1,x_2}(x_1, x_2)$. This is obtained from the joint histogram shown in Fig. 3.2(b). It is a two-dimensional "bar graph". The volumes of the bars correspond to the probabilities of x_1 and x_2 "hitting" or falling within the corresponding quantum ranges of x_1 and x_2 at times t_1 and t_2, respectively. The two-dimensional probability density function, shown in Fig.3.2(c), is obtained as the limit of the two-dimensional histogram as the quantum box size is made smaller and smaller and approaches zero in the limit.

Taking three samples in sequence, x_1, x_2, and x_3, at times t_1, t_2, and t_3, one could construct a histogram and from it a three-dimensional PDF, represented by $f_{x_1,x_2,x_3}(x_1, x_2, x_3)$. Similarly, histograms and PDFs are obtainable from multiple

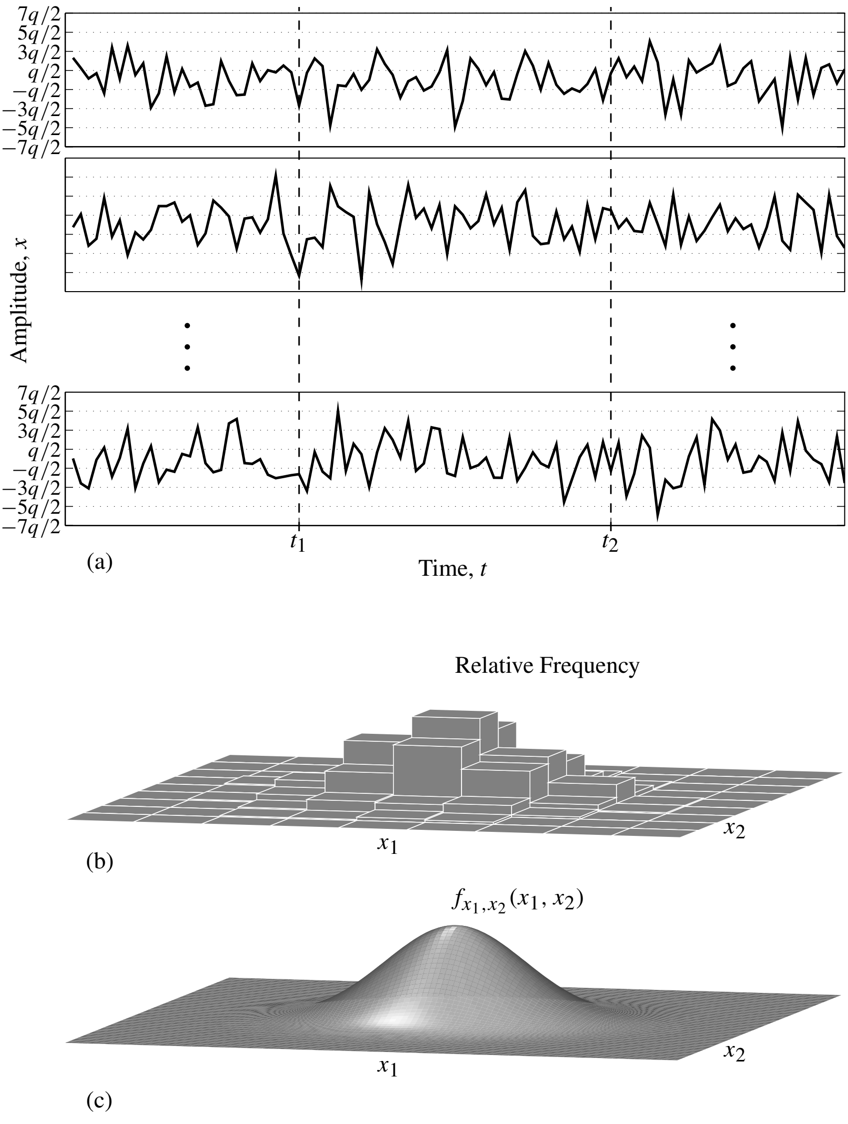

Figure 3.2 Derivation of a two-dimensional histogram and PDF: (a) an ensemble of random time functions; (b) a histogram of x_1, x_2; (c) the PDF of x_1, x_2.

samples of x, taken sequentially in time. In like manner, multidimensional histograms and PDFs can be obtained by simultaneously sampling a set of related random variables.

For two variables x_1 and x_2, the probability of a joint event is

$$f_{x_1,x_2}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 ,$$

corresponding to the first variable being within the range $x_1 \pm 0.5\mathrm{d}x_1$ while the second variable is within the range $x_2 \pm 0.5\mathrm{d}x_2$. If the two variables are statistically independent, the probability of the joint event is the product of the probabilities of the individual events, i.e., $f_{x_1}(x_1)\mathrm{d}x_1 \cdot f_{x_2}(x_2)\mathrm{d}x_2$. Therefore,

$$\begin{aligned} f_{x_1,x_2}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 &= f_{x_1}(x_1)\mathrm{d}x_1 \cdot f_{x_2}(x_2)\mathrm{d}x_2 , \text{ or} \\ f_{x_1,x_2}(x_1, x_2) &= f_{x_1}(x_1) f_{x_2}(x_2) . \end{aligned} \tag{3.20}$$

Equation (3.20) is a necessary and sufficient condition for statistical independence.

In general, the joint PDF is a complete statistical description of the variables and of their statistical connection. We should note that the total volume under this PDF is unity, i.e.,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = 1 . \tag{3.21}$$

The PDF of x_2 may be obtained by integration of the joint PDF over all values of x_1, as is done in Eq. (3.22). This yields $f_{x_2}(x_2)$, a "marginal PDF" of $f_{x_1,x_2}(x_1, x_2)$:

$$\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2)\,\mathrm{d}x_1 = f_{x_2}(x_2) . \tag{3.22}$$

That $f_{x_2}(x_2)$ results from such an integration can be justified by the following argument. Let the second variable take a value within the range $\pm 0.5\mathrm{d}x_2$ about a certain value x_2, while the first variable takes on, in general, one of an infinite range of values. The probability that this will happen is the sum of the probabilities of the mutually exclusive events corresponding to the first variable taking on all possible values, while the second variable lies in the range $x_2 \pm 0.5\mathrm{d}x_2$. Thus,

$$f_{x_2}(x_2)\mathrm{d}x_2 = \mathrm{d}x_2 \int_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2)\,\mathrm{d}x_1 . \tag{3.23}$$

Since the area under $f_{x_2}(x_2)$ must be unity, integrating both sides of (3.23) gives

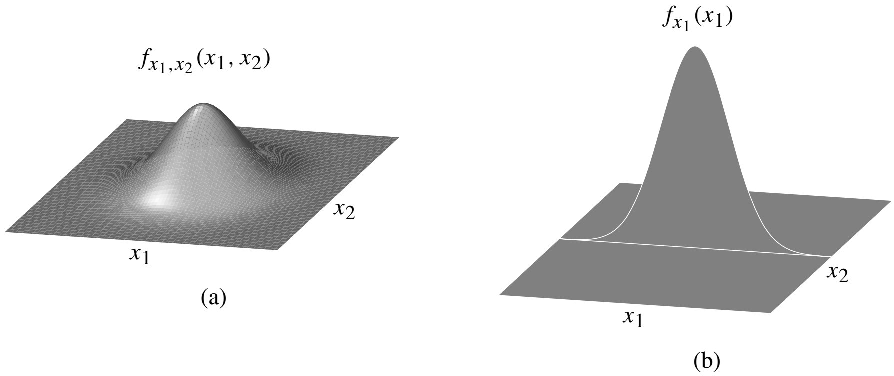

Figure 3.3 Joint PDF of x_1 and x_2, and integral with respect to x_2: (a) two-dimensional PDF; (b) one-dimensional marginal PDF.

$$\int_{-\infty}^{\infty} f_{x_2}(x_2)\,\mathrm{d}x_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = 1\,. \tag{3.24}$$

The same reasoning can be applied to give the other marginal PDF of $f_{x_1,x_2}(x_1, x_2)$:

$$\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2)\,\mathrm{d}x_2 = f_{x_1}(x_1)\,. \tag{3.25}$$

A sketch of a two-dimensional PDF is given in Fig. 3.3(a). The marginal density $f_{x_1}(x_1)$ is shown in Fig. 3.3(b). It can be thought of as the result of collapsing the volume of $f_{x_1,x_2}(x_1, x_2)$ onto a vertical plane through the f and x_2 axes.

A different view of the two-dimensional PDF $f_{x_1,x_2}(x_1, x_2)$ is shown in Fig. 3.4(a). In Fig. 3.4(b), a top view looking down shows contours of constant probability density. A vertical plane parallel to the x_2-axis cuts the surface and gives rise to a section proportional to what is called a "conditional PDF". The conditional PDF is the probability density of one of the variables given that the other variable has taken a certain value. This conditional PDF must have a unit area. The vertical section in Fig. 3.4(a) when normalized so that its area is unity (its actual area is $\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2)\,\mathrm{d}x_1 = f_{x_2}(x_2)$), gives the PDF of the variable x_1, given that x_2 has a certain value. This conditional PDF is indicated by $f_{x_1|x_2}(x_1|x_2)$. Other vertical

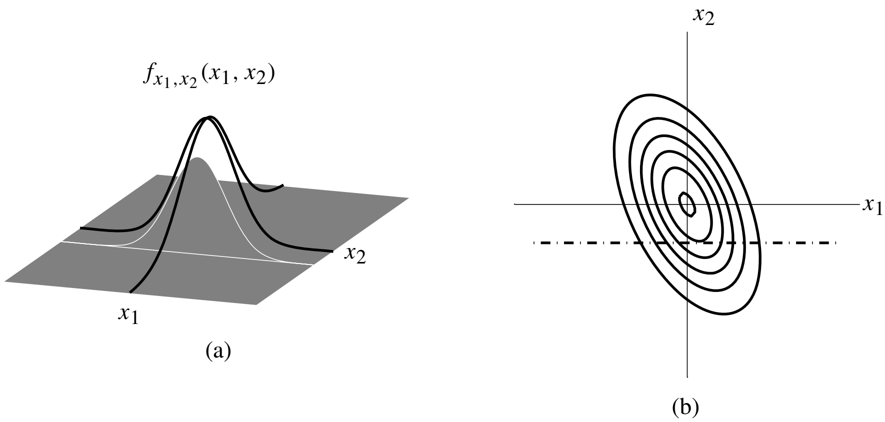

Figure 3.4 A two-dimensional PDF: (a) perspective view; (b) contour lines.

sections parallel to the x_1 axis when normalized give rise to the conditional densities $f_{x_2|x_1}(x_2|x_1)$. A formal expression of this normalization gives the conditional density

$$f_{x_2|x_1}(x_2|x_1) = \frac{f_{x_1,x_2}(x_1, x_2)}{\int\limits_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2)\,\mathrm{d}x_2} = \frac{f_{x_1,x_2}(x_1, x_2)}{f_{x_1}(x_1)}\,. \tag{3.26}$$

The other conditional density is

$$f_{x_1|x_2}(x_1|x_2) = \frac{f_{x_1,x_2}(x_1, x_2)}{f_{x_2}(x_2)}\,. \tag{3.27}$$

Going further, we may note that

$$\int\limits_{-\infty}^{\infty} f_{x_2|x_1}(x_2|x_1) f_{x_1}(x_1)\,\mathrm{d}x_1 = \int\limits_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2)\,\mathrm{d}x_1 = f_{x_2}(x_2)\,. \tag{3.28}$$

To get this relation, we have used equations (3.26) and (3.23). Substituting (3.28) into (3.27), and using (3.26), we obtain

$$f_{x_1|x_2}(x_1|x_2) = \frac{f_{x_2|x_1}(x_2|x_1) f_{x_1}(x_1)}{\int\limits_{-\infty}^{\infty} f_{x_2|x_1}(x_2|x_1) f_{x_1}(x_1)\,\mathrm{d}x_1}\,. \tag{3.29}$$

This is the famous Bayes' rule, a very important relation in estimation theory. By symmetry, the other form of Bayes' rule is obtained as

$$f_{x_2|x_1}(x_2|x_1) = \frac{f_{x_1|x_2}(x_1|x_2) f_{x_2}(x_2)}{\int\limits_{-\infty}^{\infty} f_{x_1|x_2}(x_1|x_2) f_{x_2}(x_2)\,\mathrm{d}x_2}\,. \tag{3.30}$$

3.4 JOINT CHARACTERISTIC FUNCTIONS, MOMENTS, AND CORRELATION FUNCTIONS

A joint PDF (second-order PDF) of samples x_1 and x_2 is sketched in Fig. 3.5(a). The joint characteristic function is the two-dimensional Fourier transform of the PDF given by

$$\begin{aligned}\Phi_{x_1,x_2}(u_1, u_2) &= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2) e^{j(u_1x_1+u_2x_2)}\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= \mathrm{E}\{e^{ju_1x_1+ju_2x_2}\}\,. \end{aligned} \tag{3.31}$$

The conjugate symmetry holds, similarly to the one-dimensional case (see Eq. (3.4)):

$$\Phi_{x_1,x_2}(-u_1, -u_2) = \overline{\Phi_{x_1,x_2}(u_1, u_2)}\,. \tag{3.32}$$

A two-dimensional characteristic function is sketched in Fig. 3.5(b).

The joint moments between x_1 and x_2 can be obtained by differentiation of (3.31). It is easy to verify that the (k, l)th joint moment is

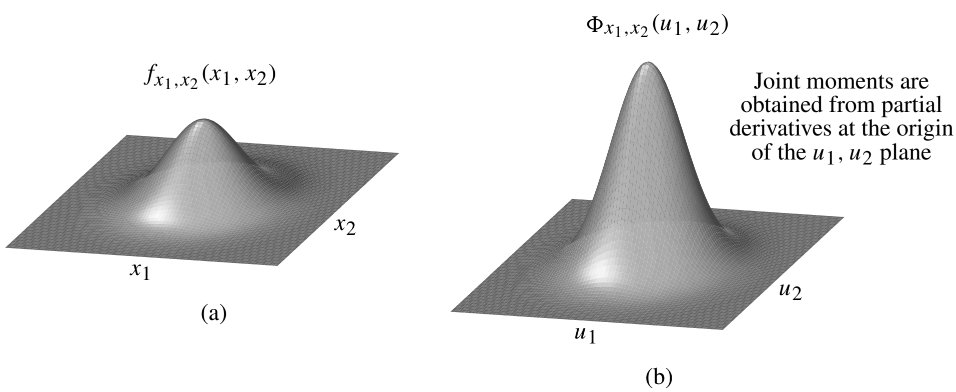

Figure 3.5 For the variables x_1 and x_2: (a) the PDF; (b) the CF.

$$
\mathrm{E}\{x_1^k x_2^l\} = \frac{1}{j^{k+l}} \frac{\partial^{k+l}\, \Phi_{x_1,x_2}(u_1,u_2)}{\partial u_1^k \partial u_2^l}\Bigg|_{\substack{u_1=0\\u_2=0}} . \tag{3.33}
$$

Thus, the second-order joint moments are related to the partial derivatives of the CF at the origin in the CF domain.

The joint PDF and CF of the variables x_1 and x_2 need further discussion. If x_1 and x_2 are independent, $\Phi_{x_1,x_2}(u_1,u_2)$ is factorable. Making use of Eq. (3.20), we have

$$
\begin{aligned}
\Phi_{x_1,x_2}(u_1,u_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2) e^{ju_1x_1} e^{ju_2x_2}\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty} f_1(x_1) e^{ju_1x_1}\,\mathrm{d}x_1 \int_{-\infty}^{\infty} f_2(x_2) e^{ju_2x_2}\,\mathrm{d}x_2 \\
&= \Phi_{x_1}(u_1)\cdot\Phi_{x_2}(u_2) .
\end{aligned} \tag{3.34}
$$

If the joint CF of x_1 and x_2 is analytic, it can be expressed in a two-dimensional Maclaurin series in terms of its derivatives at the origin. Knowledge of the moments would thereby enable calculation of the CF, which could then be inverse transformed to yield the PDF.

We will need one more relationship, concerning joint characteristic functions. If the joint characteristic function of the random variables x and ν is $\Phi_{x,\nu}(u_x,u_\nu)$, then the joint CF of x, ν, and $x' = x+\nu$ can be directly calculated from this. Consider that because of the deterministic relationship,

$$
f_{x,\nu,x'}(x,\nu,x') = \delta(x'-x-\nu) f_{x,\nu}(x,\nu) , \tag{3.35}
$$

and therefore

$$
\begin{aligned}
\Phi_{x,\nu,x'}(u_x,u_\nu,u_{x'}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x,\nu,x'}(x,\nu,x') e^{ju_xx+ju_\nu\nu+ju_{x'}x'}\,\mathrm{d}x'\,\mathrm{d}x\,\mathrm{d}\nu \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x,\nu}(x,\nu) e^{j(u_x+u_{x'})x+j(u_\nu+u_{x'})\nu}\,\mathrm{d}x\,\mathrm{d}\nu \\
&= \Phi_{x,\nu}(u_x+u_{x'},u_\nu+u_{x'}) .
\end{aligned} \tag{3.36}
$$

It is also easy to see that the marginal CFs can be obtained from the joint CFs by substituting zero for the relevant variable:

$$
\Phi_x(u_x) = \Phi_{x,y}(u_x,u_y)\Big|_{u_y=0}
$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{x,y}(x, y) e^{j(u_x x + 0 \cdot y)} \, \mathrm{d}x \, \mathrm{d}y$$
$$= \int_{-\infty}^{\infty} f_x(x) e^{j u_x x} \, \mathrm{d}x \,. \tag{3.37}$$

$\mathrm{E}\{x_1 x_2\}$ is called the *correlation* of x_1 and x_2. If this correlation is zero, the random variables are called *orthogonal*. The zero-mean versions of x_1 and x_2 are $\tilde{x}_1$ and $\tilde{x}_2$, respectively. $\mathrm{E}\{\widetilde{x}_1 \widetilde{x}_2\}$ is called the *covariance* of x_1 and x_2, $\mathrm{cov}\{x_1, x_2\}$. The variables x_1 and x_2 are *uncorrelated* if $\mathrm{cov}\{x_1, x_2\} = 0$. The *variance* of x_i is $\mathrm{E}\{\widetilde{x}_i^2\}$, and is represented by $\mathrm{var}\{x_i\}$.

If x_1 and x_2 are independent, any joint moment will be factorable, i.e.,

$$\mathrm{E}\{x_1^k x_2^l\} = \mathrm{E}\{x_1^k\}\mathrm{E}\{x_2^l\} \,. \tag{3.38}$$

This can be obtained from Eqs. (3.34) and (3.33).

A very useful measure of correlation is the correlation coefficient, defined as

$$\rho_{x_1,x_2} \stackrel{\triangle}{=} \frac{\mathrm{cov}\{x_1, x_2\}}{\sqrt{\mathrm{var}\{x_1\}\,\mathrm{var}\{x_2\}}} \,. \tag{3.39}$$

This coefficient is dimensionless, and will always have a value in the range $-1 \le \rho \le 1$. For example, with $x_2 = x_1$, the correlation between x_1 and x_2 is perfect, and ρ will be 1. With $x_2 = -x_1$, ρ will have a value of -1. It should be noted that when $\rho = 0$, x_1 and x_2 are uncorrelated but not necessarily independent. When x_1 and x_2 are independent, they are also uncorrelated, and if, in addition, either or both have zero mean, $\mathrm{E}\{x_1 x_2\} = 0$, that is, they are orthogonal. If $x(t_1)$ and $x(t_2)$ are adjacent samples of the same process, $R_{xx}(t_1, t_2) = \mathrm{E}\{x(t_1)x(t_2)\}$ is called the autocorrelation.

If x_1 and x_2 are taken from a stationary ensemble, then the joint PDF depends only on the time separation between x_1 and x_2, not on when x_1 and x_2 were taken. The autocorrelation between x_1 and x_2 then depends on their time separation (let us call this τ). The "autocorrelation function" is represented by

$$R_{xx}(\tau) = \mathrm{E}\{x(t)x(t + \tau)\} \,. \tag{3.40}$$

If x and y are samples of two different stationary variables taken at times t_1 and t_2 respectively, and t_1 and t_2 are separated by τ, the "crosscorrelation function" between x and y is represented by

$$R_{xy}(\tau) = \mathrm{E}\{x(t)y(t + \tau)\} \,. \tag{3.41}$$

In many practical cases, ensemble averaging cannot be performed, e. g. because an ensemble of sample functions is not available. We can however do averaging along the time axis. The result of time averaging often equals the result of ensemble averaging. In this case, we say that the random process is *ergodic*:

$$
\begin{aligned}
\mathrm{E}\{x\} &= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)\,\mathrm{d}t \\
\mathrm{E}\{x(t)x(t+\tau)\} &= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau)\,\mathrm{d}t \\
&\text{etc.}
\end{aligned}
\tag{3.42}
$$

3.5 FIRST-ORDER STATISTICAL DESCRIPTION OF THE EFFECTS OF MEMORYLESS OPERATIONS ON SIGNALS

How statistics "propagate" in memoryless systems is easy to visualize. An example of a memoryless system, a gain of 2, is shown in Fig. 3.6(a). The PDF of the input $f_x(x)$ is mapped into the PDF of the output $f_g(g)$ as

$$f_g(g) = \frac{1}{2} f_x\left(\frac{1}{2} g\right). \tag{3.43}$$

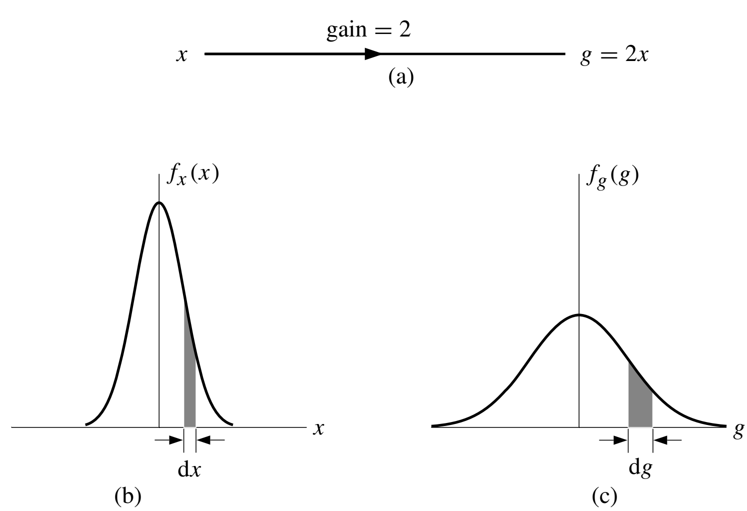

Figure 3.6 Mapping a PDF through a memoryless linear amplifier with a gain of 2: (a) a gain of 2; (b) the PDF of x; (c) the PDF of g.

The output PDF is halved in amplitude and doubled in width, relative to the input PDF. This corresponds to the output variable covering twice as much dynamic range and spending only half the time in corresponding amplitude ranges. The shaded area of $f_x(x)$, shown in Fig. 3.6(b), is the probability of the input being within the x-range indicated. The corresponding shaded area of $f_g(g)$, shown in Fig. 3.6(c), has twice as much base and hence 1/2 the amplitude:

$$|\mathrm{d}g| = 2|\mathrm{d}x|\,, \qquad \text{and} \qquad f_x(x)|\mathrm{d}x| = f_g(g)|\mathrm{d}g|\,. \tag{3.44}$$

The absolute value signs are not needed here, but would be necessary if the gain were negative. It should be noted that probability densities are always nonnegative. From Eq. (3.44) a useful expression is obtained:

$$f_g(g) = f_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}g}\right|\,. \tag{3.45}$$

In the characteristic function domain, we can relate $\Phi_x(u)$ and $\Phi_g(u)$ for the amplifier of gain 2. Making use of Eq. (3.43), and writing the CFs as Fourier transforms of the PDFs, we conclude that

$$\Phi_g(u) = \Phi_x(2u)\,. \tag{3.46}$$

Because of the Fourier transform relation between the CF and the PDF, "narrow" in the PDF domain means "wide" in the characteristic function domain and vice versa. Regarding (3.46), $\Phi_g(u)$ and $\Phi_x(u)$ have values of 1 at $u = 0$. This corresponds to both $f_x(x)$ and $f_g(g)$ having areas of unity.

An intuitive way to think of a characteristic function is as the expected value (see Eq. (3.12))

$$\Phi(u) = \mathrm{E}\left\{e^{ju(\text{variable whose CF is desired})}\right\}\,. \tag{3.47}$$

For example, the characteristic function of x is

$$\Phi_x(u) = \int_{-\infty}^{\infty} f_x(x)e^{jux}\,\mathrm{d}x\,, \tag{3.48}$$

and the characteristic function of $g = 2x$ is

$$\Phi_g(u) = \int_{-\infty}^{\infty} f_x(x)e^{ju2x}\,\mathrm{d}x = \Phi_x(2u)\,. \tag{3.49}$$

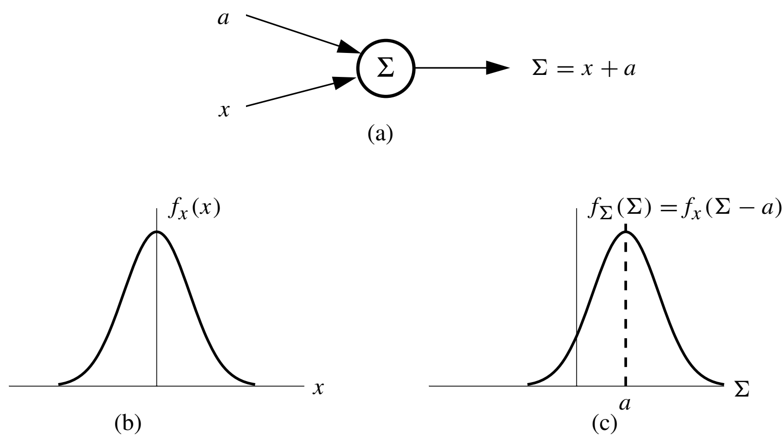

Figure 3.7 Effects of addition of a constant to the random variable x: (a) addition of a to the random variable x; (b) the PDF of x; (c) the PDF of $x + a$.

This agrees with Eq. (3.46). For a general function g, the CF is

$$\Phi_g(u) = \int_{-\infty}^{\infty} f_g(g) e^{jug} \, \mathrm{d}g = \int_{-\infty}^{\infty} f_x(x) e^{jug(x)} \, \mathrm{d}x \, . \tag{3.50}$$

The effects of the *addition of a constant* to $x(t)$ have already been considered. We will look at this again from another viewpoint. Fig. 3.7 shows the shift in the PDF of a signal whose mean value is increased by a. Note that $\Sigma = x + a$. Now, by inspection,

$$\Phi_\Sigma(u) = e^{jua} \, \Phi_x(u) \, . \tag{3.51}$$

Addition of a constant shifts the PDF like a delay, and introduces a linear phase shift to the CF. The function $\Phi_\Sigma(u)$ can also be derived in a formal way however:

$$\Phi_\Sigma(u) = \int_{-\infty}^{\infty} f_x(x) e^{ju(x+a)} \, \mathrm{d}x = e^{jua} \, \Phi_x(u) \, . \tag{3.52}$$

This agrees with Eq. (3.15).

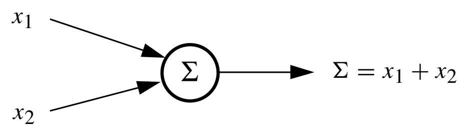

Figure 3.8 Addition of two random variables.

3.6 ADDITION OF RANDOM VARIABLES AND OTHER FUNCTIONS OF RANDOM VARIABLES

We turn next to the problem of how to find the PDF of the sum of two random variables, given their joint PDF, $f_{x_1,x_2}(x_1, x_2)$. The random variables x_1 and x_2 are added in Fig. 3.8. The PDF of the sum can be obtained by first finding its CF. Given that $\Sigma = x_1 + x_2$,

$$\begin{aligned}\Phi_\Sigma(u) &= \int_{-\infty}^{\infty} f_\Sigma(\Sigma) e^{ju\Sigma}\,\mathrm{d}\Sigma = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2) e^{ju\Sigma}\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1, x_2) e^{ju(x_1+x_2)}\,\mathrm{d}x_1\,\mathrm{d}x_2\,. \end{aligned} \tag{3.53}$$

Thus, the CF of Σ is obtained by taking the expectation of $e^{ju\Sigma}$. The PDF of the sum may therefore be gotten by inverse Fourier transformation of the CF. The moments of the sum may be gotten by differentiation of the CF at $u = 0$.

If x_1 and x_2 are independent,

$$\Phi_\Sigma(u) = \int_{-\infty}^{\infty} f_{x_1}(x_1) e^{jux_1}\,\mathrm{d}x_1 \int_{-\infty}^{\infty} f_{x_2}(x_2) e^{jux_2}\,\mathrm{d}x_2 = \Phi_{x_1}(u)\,\Phi_{x_2}(u)\,. \tag{3.54}$$

Thus, the characteristic function of the sum of two independent variables is the product of their CFs, and the PDF of the sum is the convolution of the two PDFs,

$$f_\Sigma(z) = f_{x_1}(z) \star f_{x_2}(z)\,. \tag{3.55}$$

This can be obtained as

$$f_\Sigma(z) = \int_{-\infty}^{\infty} f_{x_1}(z - x) f_{x_2}(x)\,\mathrm{d}x\,, \tag{3.56}$$

or equivalently as

$$f_\Sigma(z) = \int_{-\infty}^{\infty} f_{x_1}(x) f_{x_2}(z-x)\,\mathrm{d}x\,. \tag{3.57}$$

If x_1 and x_2 were multiplied together, the CF of their product $\Phi_\pi(u)$ would be:

$$\begin{aligned}\Phi_\pi(u) &= \int_{-\infty}^{\infty} f_\pi(\pi) e^{ju\pi}\,\mathrm{d}\pi = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1,x_2) e^{ju\pi}\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1,x_2) e^{ju(x_1 x_2)}\,\mathrm{d}x_1\,\mathrm{d}x_2\,. \end{aligned} \tag{3.58}$$

Even if x_1 and x_2 were statistically independent, the CF of $\Phi_\pi(u)$ would not be factorable. The double integral in (3.58) would not be separable, even when the joint PDF is separable. Consequently, the CF of the product of two statistically independent variables cannot be expressed as a simple function of the two PDFs. The joint PDF must be used. To get the PDF of the product, one obtains it as the inverse transform of $\Phi_\pi(u)$.

The CF of a general function $g(x_1, x_2)$ of two dependent variables x_1 and x_2 is given by Eq. (3.59). Notice the similarity between this form and the one-variable version Eq. (3.50).

$$\Phi_g(u) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1,x_2}(x_1,x_2) e^{jug(x_1,x_2)}\,\mathrm{d}x_1\,\mathrm{d}x_2\,. \tag{3.59}$$

The PDF of this function could be obtained by inverse transformation of $\Phi_g(u)$.

3.7 THE BINOMIAL PROBABILITY DENSITY FUNCTION

A good application for the characteristic function is the analysis of a *one-dimensional random walk*. The classical picture is that of a drunken person who can take a step forward or a step back with probabilities p and q, respectively. $p + q = 1$. The question is, what is the PDF of the person's position after n steps?

Figure 3.9(a) shows a line that the person steps along. The PDF of a single step is shown in Fig. 3.9(b).

Figure 3.10 shows a summation of n binary variables, each corresponding to a single step in the random walk. The figure shows the person's position formed as a sum of n steps. The values of the xs in Fig. 3.10 are the random positional increments on a given n-step random walk.

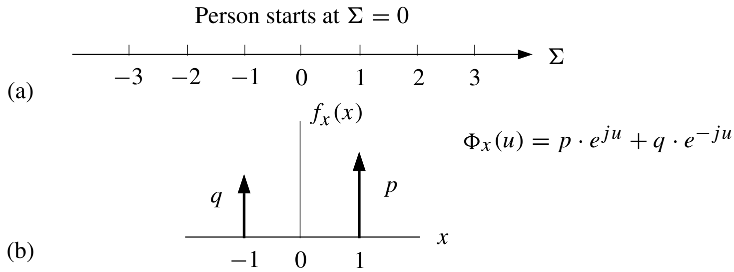

Figure 3.9 One-dimensional random walk: (a) a sum-line; (b) the PDF of a single step.

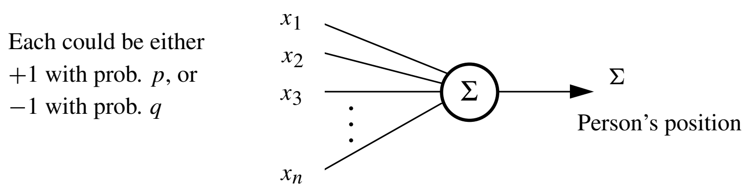

Figure 3.10 Addition of steps to get the person's position.

It should be evident that we are interested in the PDF resulting from the addition of n statistically independent random variables, each having the PDF shown in Fig. 3.9(b). The CF of a given variable x in Fig. 3.10 is the transform of two impulses, one "advanced" and one "delayed," as shown in Fig. 3.9. It follows that the CF of the sum, the product of the individual CFs is

$$\begin{aligned}\Phi_\Sigma(u) &= (pe^{ju} + qe^{-ju})^n \\ &= p^n e^{jnu} + np^{n-1}qe^{j(n-2)u} + \frac{n(n-1)}{2!}p^{n-2}q^2e^{j(n-4)u} \ldots \\ &\qquad \ldots + npq^{n-1}e^{-j(n-2)u} + q^n e^{-jnu}\,. \end{aligned} \tag{3.60}$$

Notice that at $u = 0$, $\Phi_\Sigma(0) = (p+q)^n = 1$, as all physical CFs must.

For the special case, where $p = q = 1/2$, (3.60) becomes

$$\begin{aligned}\Phi_\Sigma(u) &= \Big((1/2)e^{ju} + (1/2)e^{-ju}\Big)^n \\ &= (1/2)^n e^{jnu} + n(1/2)^{n-1}(1/2)e^{j(n-2)u} \\ &\qquad + \frac{n(n-1)}{2!}(1/2)^{n-2}(1/2)^2 e^{j(n-4)u} + \ldots \\ &\qquad \ldots + n(1/2)(1/2)^{n-1}e^{-j(n-2)u} + (1/2)^n e^{-jnu}\end{aligned}$$

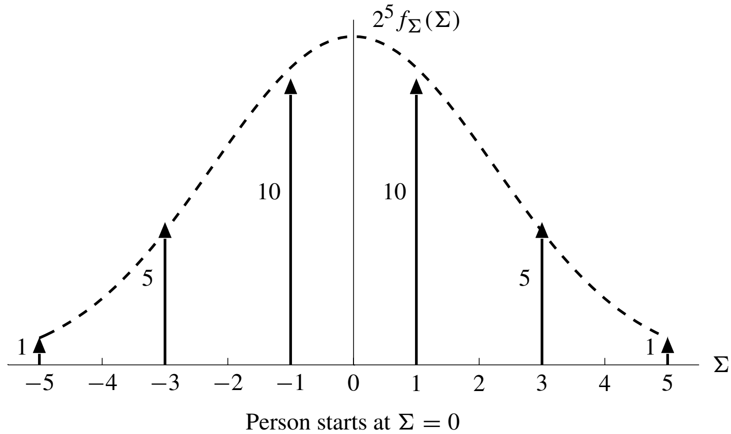

Figure 3.11 The binomial distribution, n=5, p=q=1/2.

$$= (1/2)^n \bigg(e^{jnu} + ne^{j(n-2)u} + \frac{n(n-1)}{2!} e^{j(n-4)u} + \dots \\ \dots + ne^{-j(n-2)u} + e^{-jnu} \bigg) . \tag{3.61}$$

When taking the inverse Fourier transform, this CF leads directly to the binomial PDF, plotted in Fig. 3.11 for $n = 5$ jumps. The areas of the impulses correspond to the binomial coefficients.

If p and q are not equal, the PDF drifts (develops a bias) as n increases. In any event, the PDF spreads as n is increased. When p and q are not equal, we have a propagating spreading "packet of probability." All this can be seen from the general $\Phi_\Sigma(u)$ and its inverse Fourier transform.

3.8 THE CENTRAL LIMIT THEOREM

The PDF of the sum of a large number of random, statistically independent variables of comparable variances almost invariably has a "bell-shaped" character, without regard to the shapes of the PDFs of the individual variables. The PDF of the sum closely approximates the normal or Gaussian PDF (see also Appendix F), shown for zero mean in Fig. 3.12(a). The CF (which is also Gaussian) is shown in Fig. 3.12(b). Thus, the PDF of the sum of many independent random variables can usually be approximated by a Gaussian PDF having the right mean and variance. The variance of the Gaussian PDF is represented by σ^2 which is set equal to the sum of the variances of the independent variables. The mean of the sum is the sum of their means.