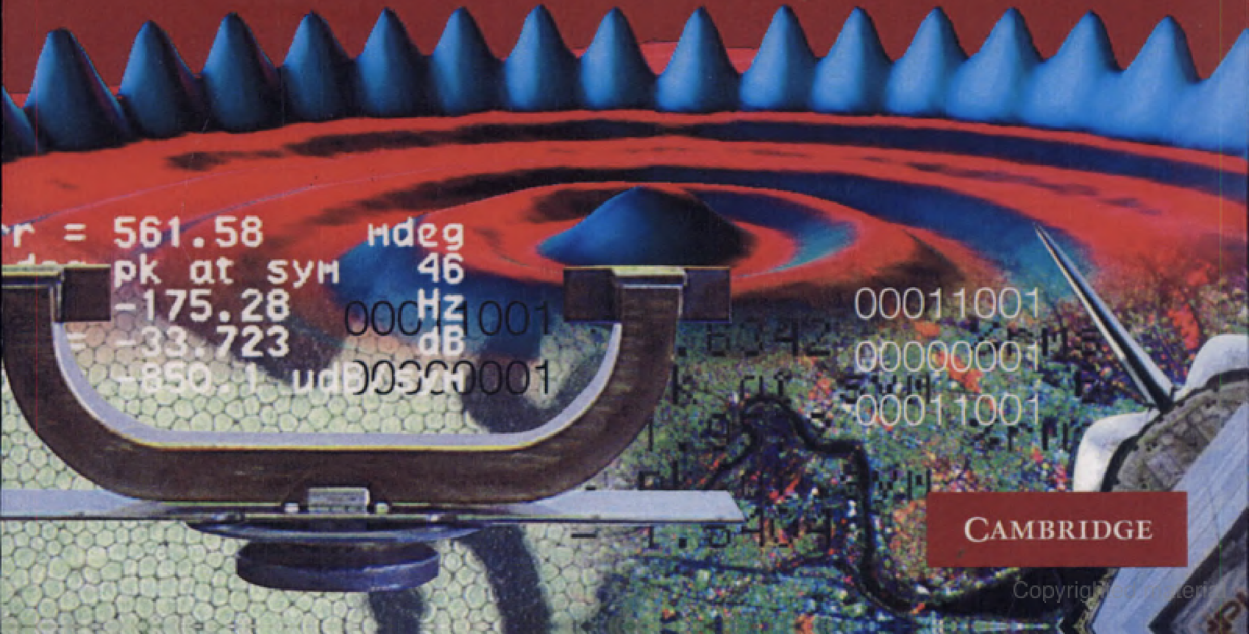
The Physics of Information Technology

NEIL GERSHENFELD



CAMBRIDGE

The Physics of Information Technology

Neil Gershenfeld

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2000
Reprinted 2002

Printed in the United Kingdom at the University Press, Cambridge

Typeset in Monotype Ehrhardt 10½/13pt, in LaTeX 2ε [EPC]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

ISBN 0 521 58044 7 hardback

Contents

Preface

How does the bandwidth of a telephone line relate to the bit rate that can be sent through it? Modems keep getting faster; how quickly can they operate? These sensible questions have unexpectedly profound answers. At MIT, I've been asked them by people ranging from undergrads to faculty. A good engineer might know about coding theory and the concept of channel capacity, but not understand the origin of the noise that limits the capacity. Conversely, a physicist might use the fluctuation–dissipation theorem to explain why resistors are necessarily noisy, but know nothing of information theory. And the computer scientist sending data over the phone line might not understand either side. The most interesting aspects of this problem can easily be missed among these poles. I've found this pattern to recur over and over: people may not appreciate the useful applications of fundamental results in the devices they use, or the deep implications of their practical knowledge, and may not have a good sense of how their formal academic training can relate to their personal passions.

The familiar computing and communications devices that we use to manipulate information operate near many remarkable physical limits. A handheld GPS receiver applies both special- and general-relativistic corrections to its timing measurements of signals from atomic clocks in satellites in order to maintain the system's global 1 ns accuracy. The head in a high-capacity disk drive flies within a single mean-free-path of an air molecule above the platter, and so the aerodynamic design problem can no longer be solved by modeling the airflow with continuum partial differential equations. This kind of tremendous ingenuity has gone into finding practical solutions to what had appeared to be impossible technological problems. However, the exponential improvements that we've come to rely on, such as processor speeds doubling every few years, must stop when current scaling trends run into basic physical limits. Circuits cannot have wires smaller than atoms, signals faster than light, or charge carriers less than an electron. Given such constraints, a CMOS chip that can perform 10^9 floating-point operations per second (a gigaflop) is feasible, but 10^{12} (a teraflop) is unlikely. Understanding these kinds of systems requires equal familiarity with fundamental physics and with very practical engineering. Because this kind of background is hard to develop given the traditional split between basic and applied science, it's easy for students (and practitioners) to run into either the Scylla of uncritically accepting the received wisdom of past practice, or the Charybdis of enthusiastically pursuing impossible alternatives.

This book grew out of lecture notes for a course that I've developed at MIT's Media Lab. The goal is to review basic physical governing equations in a number of areas relevant to information technology, and then work up through device mechanisms to a

1 Introduction

Why does computation require energy?

Because there must be some irreversibility to ensure that calculations go forward (from inputs to outputs) and not in reverse, and because logical erasure necessarily implies dissipation because of the compression of phase-space.

What is a quantum computer?

One that operates on quantum bits that can be in a superposition of many different states simultaneously and that maintain a connection (called entanglement) following an interaction. These properties change the computational order of many important problems, such as reducing factoring from requiring a time that is exponentional to polynomial in the number of bits.

What limits the bit density for semiconductor memory?

Lithography (constrained by the wavelength used to pattern a memory cell, and the resulting yield), electromigration (when too few atoms are used in a wire they move in response to currents), and capacitance (when too few electrons are used, the fluctuation in their number becomes significant).

What limits the bit density in a typical hard disk?

Magnetic domain wall energies, and the head height.

What limits the bit density for optical storage?

The diffraction limit for focusing light, which is proportional to the wavelength.

Why are twisted pairs twisted, and coaxial cables coaxial?

To reduce the generation of unwanted radiation and the sensitivity to interference, and to effectively guide the signal. Twisted pairs are best at low frequencies, and coaxial cables at high frequencies.

Where does electronic noise come from, and how does it limit data rates?

Thermodynamic fluctuations, defect scattering, and finite-size statistics. The capacity of a communications channel grows as the logarithm of the ratio of the energy in the signal and the noise.

What is a liquid crystal, and how does it modulate light?

It is a material that maintains long-range orientational ordering without translational ordering. Under an applied field it is able to rotate the direction of polarization of light, thereby modulating the intensity of the light if the material is enclosed between polarizers.

These questions are examples of the many ways in which familiar devices that detect, transmit, process, store, and deliver information operate surprisingly near fundamental physical limits. The goal of this book is to explore how such devices function, how they can be used, what the limits on their performance are, and how they might be improved. This will require developing familiarity with the physical governing equations for a range of types of behavior, and with the mathematical tools necessary to manipulate these equations. One important aim is to equip the reader to work out quantitative answers to questions such as these.

A note about pedagogy: reading about physics is as satisfying as reading about food or exercise. It can be useful, but there is no substitute for experience solving problems. Each chapter has problems that apply and develop the preceeding ideas, ranging from trivial calculations to open research questions. Since another goal of this book is to help develop problem-solving skills, consulting the supplied answers before a problem is attempted is entirely counter-productive because the real problems that will come after this book don't come with such handy answers.

And a note about epistemology: it is important to keep in mind the distinction between truth and models. I will be describing models for a variety of types of behavior; these are the product of both experimental observations and theoretical inferences. A good model should compactly explain what you already know and allow you to predict new things that you did not know, but it does not necessarily contain any guide to an underlying "truth." Some physicists believe that there is an ultimate "correct" answer that these models are approaching, and some violently disagree, yet all agree on the usefulness of the current set of models and on how to manipulate them. *Truth* and *Meaning* are concepts that one may choose to associate with these models, but their presence or absence does not affect the models' use. At most, they do guide what you choose to think about. This distinction is very important because, when faced with unexpected claims or results, there is a recurring danger of seeing particular models as privileged correct answers rather than being open-minded about judging evidence on its merits. The history of science is littered with conflicts arising from prior beliefs that were stronger than experimental observations.

2 Interactions, Units, and Magnitudes

Modern information technology operates over a spectacular range of scales; bits from a memory cell with a size of 10^{-7} meters might be sent 10^7 meters to a geosynchronous satellite. It is important to be comfortable with the orders of magnitudes and associated interaction mechanisms that are useful in practice. Our first task will be to review the definitions of important units, then survey the types of forces, and finally look at typical numbers in various regimes.

2.1 UNITS

Many powers of ten have been named because it is much easier to say something like "a femtosecond optical pulse" than "a 0.000 000 000 000 001 second optical pulse" when referring to typical phenomena at that scale (a cycle of light takes on the order of a femtosecond). The dizzying growth of our ability to work with large and small systems pushes the bounds of this nomenclature; data from terabyte storage systems is read out into femtofarad memory cells. It is well worth memorizing the prefixes in Table 2.1.

Physical quantities must of course be measured in a system of units; there are many alternatives that are matched to different regimes and applications. Because of their interrelationships it is necessary only to define a small number of fundamental quantities to be able to derive all of the other ones. The choice of which fundamental definitions to use changes over time to reflect technological progress; once atomic clocks made it possible to measure time with great *precision* (small variance) and *accuracy* (small bias), it became more reliable to define the meter in terms of time and the speed of light rather than a reference bar kept at the Bureau International des Poids et Mesures (BIPM, `http://www.bipm.fr`) in Sevres, France. The kilogram is still defined in terms of a platinum–iridium cylinder held at BIPM instead of a fundamental physical process, a source of great frustration in the metrology community. Aside from the difficulty in duplicating it, the accumulation of contaminants on the surface increases the mass by about 1 part in 10^9 per year, requiring that it be measured only after a special cleaning procedure [Girard, 1994].

The most common set of base defined quantities in use is the *Système International d'Unités* (*SI*) [BIPM, 1998]:

length: *meter* (m)

The meter is the length of path traveled by light in vacuum during a time interval of 1/299 792 458 of a second.

Table 2.1. *Orders of magnitude.*

Magnitude	Prefix	Symbol	Magnitude	Prefix	Symbol
10^{-24}	yocto	y	10^{24}	yotto	Y
10^{-21}	zepto	z	10^{21}	zetta	Z
10^{-18}	atto	a	10^{18}	exa	E
10^{-15}	femto	f	10^{15}	peta	P
10^{-12}	pico	p	10^{12}	tera	T
10^{-9}	nano	n	10^{9}	giga	G
10^{-6}	micro	μ	10^{6}	mega	M
10^{-3}	milli	m	10^{3}	kilo	k
10^{-2}	centi	c	10^{2}	hecto	h
10^{-1}	deci	d	10^{1}	deka	da

mass: *kilogram* (kg)

The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram.

time: *second* (s)

The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom.

current: *ampere* (A)

The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 meter apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newtons per meter of length. (See Problem 5.4.)

temperature: *kelvin* (K)

The kelvin, the unit of thermodynamic temperature, is the fraction of 1/273.16 of the thermodynamic temperature of the triple point of water. (Temperatures in degrees Celsius are equal to temperatures in kelvin + 273.15. The triple point is the temperature and pressure at which the liquid, solid, and gas phases of water co-exist. It is fixed at 0.01 °C, and provides a more reliable reference than the original centigrade definition of 0 °C as the freezing point of water at atmospheric pressure.)

quantity: *mole* (mol)

The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kg of carbon 12 (i.e., Avogadro's constant $6.022\ldots\times10^{23}$).

intensity: *candela* (cd)

The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in the direction of 1/683 watts per steradian. (The frequency corresponds to the wavelength of 555 nm where the eye is most sensitive, the factor of 683 comes from matching an earlier definition based on the emission from solidifying platinum, and a steradian is the solid angle subtended by a unit

Table 2.2. *Selected conversion factors.*

1 dyne $(\mathrm{gm \cdot cm \cdot s^{-2}})$	=	1×10^{-5} N
1 erg $(\mathrm{gm \cdot cm^2 \cdot s^{-2}})$	=	1×10^{-7} J
1 horsepower (hp)	=	745.7 W
1 atmosphere (atm)	=	101325 Pa
1 ton (short)	=	2000 pounds
	=	907.18474 kg
1 electron volt (eV)	=	$1.602176462 \times 10^{-19}$ J
1 amu	=	$1.66053873 \times 10^{-27}$ kg
1 ångstrom (Å)	=	1×10^{-10} m
1 fermi (fm)	=	1×10^{-15} m
1 parsec (pc)	=	3.085678×10^{16} m
1 mile (mi)	=	1609.344 m
1 foot (ft)	=	0.3048 m
1 inch (in)	=	0.0254 m
1 liter (L)	=	0.001 $\mathrm{m^3}$
1 pound (lb)	=	0.45359237 kg
1 pound-force (lbf)	=	4.44822 N

resistance: *ohm* Ω $(\mathrm{m^2 \cdot kg \cdot s^{-3} \cdot A^{-2}})$

> The ohm is the electric resistance between two points of a conductor when a constant difference of potential of 1 volt, applied between these two points, produces in this conductor a current of 1 ampere. (These derivative definitions of the volt and ohm have more recently been replaced by fundamental ones fixing them in terms of the voltage across a *Josephson junction* and the resistance steps in the *quantum Hall effect* [Zimmerman, 1998], and capacitance may be defined by counting electrons on a *Single-Electron Tunneling* (*SET*) device [Keller *et al.*, 1999].)

It is important to pay attention to the units in these definitions. Many errors in calculations can be caught by making sure that the final units are correct, and it can be possible to make a rough estimate of an answer to a problem simply by collecting relevant terms with the right units (this is the subject of *dimensional analysis*). Electromagnetic units are particularly confusing; we will consider them in more detail in Chapter 5. The SI system is also called *MKS* because it bases its units on the meter, the kilogram, and the second. For some problems it will be more convenient to use *CGS* units (based on the centimeter, the gram, and the second); MKS is more common in engineering and CGS in physics. A number of other units have been defined by characteristic features or by historical practice; some that will be useful later are given in Table 2.2.

It's often more relevant to know the value on one quantity relative to another one, rather than the value itself. The ratio of two values X_1 and X_2, measured in *decibels* (*dB*), is defined to be

$$\mathrm{dB} = 20 \log_{10} \frac{X_1}{X_2} \quad . \tag{2.1}$$

If the *power* (energy per time) in two signals is P_1 and P_2, then

$$\mathrm{dB} = 10 \log_{10} \frac{P_1}{P_2} \quad . \tag{2.2}$$

Table 2.3. *Selected fundamental constants.*

gravitational constant (G)	=	$6.673(10)\times10^{-11}\ \mathrm{m^3\cdot kg^{-1}\cdot s^{-2}}$
speed of light (c)	=	$2.99792458\times10^{8}\ \mathrm{m/s}$
elementary charge (e)	=	$1.602176462(63)\times10^{-19}\ \mathrm{C}$
Boltzmann constant (k)	=	$1.3806503(24)\times10^{-23}\ \mathrm{J/K}$
Planck constant (h)	=	$6.62606876(52)\times10^{-34}\ \mathrm{J\cdot s}$
$\hbar = h/2\pi$	=	$1.054571596(82)\times10^{-34}\ \mathrm{J\cdot s}$
Avogadro constant (N_A)	=	$6.02214199(47)\times10^{23}\ \mathrm{mol^{-1}}$
electron mass (m_e)	=	$9.10938188(72)\times10^{-31}\ \mathrm{kg}$
proton mass (m_p)	=	$1.67262158(13)\times10^{-27}\ \mathrm{kg}$
gas constant (R)	=	$8.314472(15)\ \mathrm{J\cdot mol^{-1}\cdot K^{-1}}$
vacuum permittivity (ϵ_0)	=	$10^7/(4\pi c^2) = 8.854188\ldots\times10^{-12}\ \mathrm{F/m}$
vacuum permeability (μ_0)	=	$4\pi\times10^{-7}\ \mathrm{H/m}$

This story starts with quantum mechanics, the laws that govern things that are very small. Around 1900 Max Planck was led by his inability to explain the spectrum of light from a hot oven to propose that the energy of light is quantized in units of $E = h\nu = hc/\lambda$, where ν is the frequency and λ is the wavelength; $h = 6.626\ldots\times10^{-34}\ \mathrm{J\cdot s}$ is now called *Planck's constant*. From there, in 1905 Einstein introduced the notion of massless photons as the discrete constituents of light, and in 1924 de Broglie suggested that the wavelength relationship applies to massive as well as massless particles by $\lambda = h/p$; λ is the *de Broglie wavelength*, and is a consequence of the *wave–particle duality*: all quantum particles behave as both waves and particles. An electron, or a photon, can diffract like a wave from a periodic grating, but a detector will register the arrival of individual particles. Quantum effects usually become significant when the de Broglie wavelength becomes comparable to the size of an object.

Quantum mechanical particles can be either *fermions* (such as an electron) or *bosons* (such as a photon). Fermions and bosons are as unlike as anything can be in our universe. We will later see that bosons are particles that exist in states that are symmetric under the interchange of particles, they have an integer spin quantum number, and multiple bosons can be in the same quantum state. Fermions have half-integer spin, exist in states that are antisymmetric under particle interchange, and only one fermion can be in a particular quantum state. Spin is an abstract property of a quantum particle, but it behaves just like an angular momentum (as if the particle is spinning).

Particles can interact through four possible forces: *gravitational*, *electromagnetic*, *weak*, and *strong*. The first two are familiar because they have infinite range; the latter two operate on short ranges and are associated with nuclear and subnuclear processes (the characteristic lengths are approximately 10^{-15} m for the strong force and 10^{-18} m for the weak force). The electromagnetic force is so significant because of its strength: if a quantum atom was held together by gravitational forces alone (like a miniature solar system) its size would be on the order of 10^{23} m instead of 10^{-10} m. The macroscopic forces that we feel, such as the hardness of a wall, are transmitted to us by the electromagnetic force through the electrons in our atoms interacting with electrons in the adjoining atoms in the surface, but can be much more simply described in terms of fictitious effective forces ("the wall is hard").

All forces were originally thought to be transmitted by an intervening medium, the long-sought *ether* for electromagnetic forces. We now understand that forces operate by the exchange of spin-1 gauge bosons – the photon for the electromagnetic interaction (electric and magnetic fields), the $W^{\pm}$ and Z^0 bosons for the weak interaction, and eight gluons for the strong interaction (there is not yet a successful quantum theory of gravity). *Quantum ElectroDynamics* (*QED*) is the theory of the quantum electromagnetic interaction, and *Quantum ChromoDynamics* (*QCD*) the theory of the strong interaction. The weak and electromagnetic interactions are united in the *electroweak theory*, which, along with QCD is the basis for the *Standard Model*, the current summary of our understanding of particle physics. This amalgam of experimental observations and theoretical inferences successfully predicts most observed behavior extremely accurately, with two important catches: the theory has 20 or so adjustable parameters that must be determined from experiments, and it cannot explain gravitation. *String theory* [Giveon & Kutasov, 1999], a reformulation of particle theory that starts from loops rather than points as the primitive mathematical entity, appears to address both these limitations, and so is of intense interest in the theoretical physics community even though it is still far from being able to make experimentally testable predictions.

The most fundamental massive particles that we are aware of are the *quarks* and *leptons*. There's no reason to assume that there's nothing below them (i.e., turtles all the way down); there's just not a compelling reason right now to believe that there is. Quarks and leptons appear in the scattering experiments used to study particle physics to be point-particles without internal structure, and are spin-1/2 fermions. The leptons interact through the electromagnetic and weak interactions, and come in pairs: the *electron* and the electron *neutrino* (e^-, ν_e), the *muon* and its neutrino (μ^-, ν_μ), and the *tau lepton* and its neutrino (τ^-, ν_τ). Muons and tau leptons are unstable, and therefore are seen only in accelerators, particle decay products, and cosmic rays. Because neutrinos interact only through the weak force, they can pass unhindered though a light-year of lead. But they are profoundly important for the energy balance of the universe, and if they have mass [Fukuda, 1998] it will have enormous implications for the fate of the universe. Quarks interact through the strong as well as weak and electromagnetic interactions, and they come in pairs: *up* and *down*, *charm* and *strange*, and *top* and *bottom*. These fanciful names are just labels for the underlying abstract states. The first member of each pair has charge +2/3, the second member has charge $-1/3$, and each charge flavor comes in three colors (once again, flavor and color are just descriptive names for quantum numbers).

Quarks combine to form *hadrons*; the best-known of which are the two *nucleons*. A proton comprises two ups and a down, and the neutron an up and two downs. The nucleons, along with their excited states, are called *baryons* and are fermions. Transitions between baryon states can absorb or emit spin-1 boson hadrons, called *mesons*. The size of hadrons is on the order of 10^{-15} m, and the energy difference between excited states is on the order of 10^9 electron volts (1 GeV).

The nucleus of an atom is made up of some number of protons and neutrons, bound into ground and excited states by the strong interaction. Typical nuclear sizes are on the order of 10^{-14} m, and energies for nuclear excitations are on the order of 10^6 eV (1 MeV). Atoms consist of a nucleus and electrons bound by the electromagnetic interaction; typical sizes are on the order of 1 ångstrom (Å, 10^{-10} m) and the energy difference between states is on the order of 1 eV. Notice the large difference in size between the atom and the nucleus: atoms are mostly empty space. Atoms can exist in different *isotopes* that have the same number of protons but differing numbers of neutrons, and *ions* are atoms that have had electrons removed or added.

Atoms can bond to form molecules; bond energies are on the order of 1 eV and bond lengths are on the order of 1 Å. Molecular sizes range from simple diatomic molecules up to enormous biological molecules with 10^6–10^9 atoms. Large molecules fold into complex shapes; this is called their *tertiary structure*. These shapes are responsible for the geometrical constraints in molecular interactions that govern many biochemical pathways. Predicting tertiary structure is one of the most difficult challenges in chemistry.

Macroscopic materials are described by the arrangement of their constituent atoms, and include crystals (which have complete long-range ordering), liquids and glasses (which have short-range order but little long-range order), and gases (which have little short-range order). There are also very interesting intermediate cases, such as quasiperiodic alloys called *quasicrystals* that have deterministic translational order without translational periodicity [DiVincenzo & Steinhardt, 1991], and *liquid crystals* that maintain orientational but not translational ordering [Chandrasekhar, 1992]. Most solids do not contain just a single phase; there are usually defects and boundaries between different kinds of domains.

The atomic weight of an element is equal to the number of grams equal to one mole ($N_A \approx 10^{23}$) of atoms. It is approximately equal to the number of protons and neutrons in an atom, but differs because of the mix of naturally occuring isotopes. 22.4 liters of an ideal gas at a pressure of 1 atmosphere and at room temperature will also contain a mole of atoms.

The structure of a material at more fundamental levels will be invisible and can be ignored unless energies are larger than its characteristic excitations. Although we will rarely need to descend below atomic structure, there are a number of important applications of nuclear transitions, such as nuclear power and the use of nuclear probes to characterize materials.

2.3 ORDERS OF MAGNITUDE

Understanding what is possible and what is preposterous requires being familiar with the range of meaningful numbers for each unit; the following lists include some significant ones:

Time

10^{-43} s: the Planck time (Problem 2.7)

10^{-15} s: this is the period of visible light, and a typical time scale for chemical reactions

10^{-9} s: atomic excitations and molecular rotations typically have lifetimes on the order of nanoseconds, and this is the clock cycle for the fastest computers

10^{-3} s: the shortest time difference that is consciously perceptible by people

10^{17} s: the approximate age of the observable universe

Power and Energy

1 eV: atomic excitations
10^6 eV: nuclear excitations
10^9 eV: subnuclear excitations
10^{28} eV: the Planck energy
10 W: laptop computer
100 W: workstation; human
10^4 W: car
10^5 W: supercomputer; heating and lighting a building
10^{26} W: luminosity of the sun
10^{-12} W/m^2: softest sound that can be heard
1 W/m^2: loudest sound that can be tolerated
10^7 J/kg: energy density of food
10^9 J: energy in a ton of TNT
10^{20} J: energy consumption in the US per year

Temperature

10^{-7} K: lowest temperatures obtained in solids in the laboratory
2.75 K: microwave background radiation from the Big Bang
77 K: temperature of liquid nitrogen
6000 K: temperature of the surface of the sun

Mass

10^{-27} kg: proton mass
10^{-12} kg: typical cell
10^{-5} kg: small insect
10^{16} kg: Earth's biomass
5.98×10^{24} kg: the mass of the Earth
10^{42} kg: approximate mass of the Milky Way

Length

10^{-35} m: the Planck distance
10^{-15} m: size of a proton
10^{-10} m: size of an atom
4×10^5 m: height of a Low Earth Orbit satellite above the surface
6.378×10^6 m: radius of the Earth
4×10^7 m: height of a geosynchronous satellite above the equator
10^{11} m: distance from the Earth to the Sun
10^{20} m: Milky Way radius
10^{26} m: size of the observable universe

Electromagnetic spectrum

< 0.1 Å: gamma rays
0.1–100 Å: X-rays

2.5 PROBLEMS

(2.1) (*a*) How many atoms are there in a yoctomole?

(*b*) How many seconds are there in a nanocentury? Is the value near that of any important constants?

(2.2) A large data storage system holds on the order of a terabyte. How tall would a 1 terabyte stack of floppy disks be? How does that compare to the height of a tall building?

(2.3) If all the atoms in our universe were used to write an enormous binary number, using one atom per bit, what would that number be (converted to base 10)?

(2.4) Compare the gravitational acceleration due to the mass of the Earth at its surface to that produced by a 1 kg mass at a distance of 1 m. Express their ratio in decibels.

(2.5) (*a*) Approximately estimate the chemical energy in a ton of TNT. You can assume that nitrogen is the primary component; think about what kind of energy is released in a chemical reaction, where it is stored, and how much there is.

(*b*) Estimate how much uranium would be needed to make a nuclear explosion equal to the energy in a chemical explosion in 10 000 tons of TNT (once again, think about where the energy is stored).

(*c*) Compare this to the *rest mass energy* $E = mc^2$ of that amount of material (Chapter 14), which gives the maximum amount of energy that could be liberated from it.

(2.6) (*a*) What is the approximate de Broglie wavelength of a thrown baseball?

(*b*) Of a molecule of nitrogen gas at room temperature and pressure? (This requires either the result of Section 3.4.2, or dimensional analysis.)

(*c*) What is the typical distance between the molecules in this gas?

(*d*) If the volume of the gas is kept constant as it is cooled, at what temperature does the wavelength become comparable to the distance between the molecules?

(2.7) (*a*) The potential energy of a mass m a distance r from a mass M is $-GMm/r$. What is the *escape velocity* required to climb out of that potential?

(*b*) Since nothing can travel faster than the speed of light (Chapter 14), what is the radius within which nothing can escape from the mass?

(*c*) If the rest energy of a mass M is converted into a photon, what is its wavelength?

(*d*) For what mass does its equivalent wavelength equal the size within which light cannot escape?

(*e*) What is the corresponding size?

(*f*) What is the energy?

(*g*) What is the period?

(2.8) Consider a pyramid of height H and a square base of side length L. A sphere is placed so that its center is at the center of the square at the base of the pyramid, and so that it is tangent to all of the edges of the pyramid (intersecting each edge at just one point).

(*a*) How high is the pyramid in terms of L?

(*b*) What is the volume of the space common to the sphere and the pyramid?

(This question comes from an entrance examination for humanities students at Tokyo University [*Economist*, 1993].)

and the mean square deviation from this is the *variance*:

$$\begin{aligned}\sigma^2 &= \langle (x - \langle x \rangle)^2 \rangle \\ &= \langle x^2 - 2x\langle x \rangle + \langle x \rangle^2 \rangle \\ &= \langle x^2 \rangle - \langle x \rangle^2 \quad . \end{aligned} \tag{3.4}$$

The square root of the variance is the *standard deviation* σ.

The probability distribution contains no information about the temporal properties of the observed quantity; a useful probe of this is the *autocovariance function*:

$$\langle x(t)x(t-\tau)\rangle = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t-\tau)\, dt \quad . \tag{3.5}$$

If the autocovariance is normalized by the variance then it is called the *autocorrelation function*, ranging from 1 for perfect correlation to 0 for no correlation to -1 for perfect anticorrelation. The rate at which it decays as a function of τ provides one way to determine how quickly a function is varying. In the next chapter we will introduce the *mutual information*, a much more general way to measure the relationships among variables.

3.1.2 Spectral Theorems

The *Fourier transform* of a fluctuating quantity is

$$X(f) = \lim_{T\to\infty} \int_{-T/2}^{T/2} e^{i2\pi f t} x(t)\, dt \tag{3.6}$$

and the inverse transform is

$$x(t) = \lim_{F\to\infty} \int_{-F/2}^{F/2} e^{-i2\pi f t} X(f)\, df \quad . \tag{3.7}$$

The Fourier transform is also a random variable. The *Power Spectral Density* (*PSD*) is defined in terms of the Fourier transform by taking the average value of the square magnitude of the transform

$$\begin{aligned} S(f) &= \langle |X(f)|^2 \rangle = \langle X(f) X^*(f) \rangle \\ &= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} e^{i2\pi f t} x(t)\, dt \int_{-T/2}^{T/2} e^{-i2\pi f t'} x(t')\, dt' \quad . \end{aligned} \tag{3.8}$$

X^* is the complex conjugate of X, replacing i with $-i$, and we'll assume that x is real. The power spectrum might not have a well-defined limit for a non-stationary process; *wavelets* and *Wigner functions* are examples of *time–frequency transforms* that retain both temporal and spectral information for non-stationary signals [Gershenfeld, 1999a].

The Fourier transform is defined for negative as well as positive frequencies. If the sign of the frequency is changed, the imaginary or sine component of the complex exponential changes sign while the real or cosine part does not. For a real-valued signal this means that the transform for negative frequencies is equal to the complex conjugate of the transform for positive frequencies. Since the power spectrum is used to measure energy as a function of frequency, it is usually reported as the *single-sided* power spectral density found by adding the square magnitudes of the negative- and positive-frequency components. For a real signal these are identical, and so the single-sided density differs from the *two-sided* density by an (occasionally omitted) factor of 2.

The Fourier transform can also be defined with the 2π in front,

$$
\begin{aligned}
X(\omega) &= \lim_{T\to\infty} \int_{-T/2}^{T/2} e^{i\omega t} x(t)\, dt \\
x(t) &= \lim_{\Omega\to\infty} \frac{1}{2\pi} \int_{-\Omega/2}^{\Omega/2} e^{-i\omega t} X(\omega)\, d\omega \quad .
\end{aligned}
\tag{3.9}
$$

ν measures the frequency in cycles per second; ω measures the frequency in *radians* per second (2π radians = 1 cycle). Defining the transform in terms of ν eliminates the errors that arise from forgetting to include the 2π in the inverse transform or in converting from radians to cycles per second, but it is less conventional in the literature. We will use whichever is more convenient for a problem.

The power spectrum is simply related to the autocorrelation function by the *Wiener–Khinchin Theorem*, found by taking the inverse transform of the power spectrum:

$$
\begin{aligned}
&\int_{-\infty}^{\infty} S(f) e^{-i2\pi f\tau}\, df \\
&= \int_{-\infty}^{\infty} \langle X(f) X^*(f) \rangle e^{-i2\pi f\tau}\, df \\
&= \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} \int_{-T/2}^{T/2} e^{i2\pi f t} x(t)\, dt \int_{-T/2}^{T/2} e^{-i2\pi f t'} x(t')\, dt'\; e^{-i2\pi f\tau}\, df \\
&= \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} e^{i2\pi f(t-t'-\tau)}\, df\; x(t) x(t')\, dt\, dt' \\
&= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} \delta(t-t'-\tau) x(t) x(t')\, dt\, dt' \\
&= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t) x(t-\tau)\, dt \\
&= \langle x(t) x(t-\tau) \rangle \quad ,
\end{aligned}
\tag{3.10}
$$

using the Fourier transform of a *delta function*

$$
\int_{-\infty}^{\infty} e^{i2\pi xy}\, dx = \delta(y)
$$

$$
\int_{-\infty}^{\infty} f(x) \delta(x - x_0)\, dx = f(x_0)
\tag{3.11}
$$

(one way to derive these relations is by taking the delta function to be the limit of a Gaussian with unit norm as its variance goes to zero).

The Wiener–Khinchin Theorem shows that the Fourier transform of the autocovariance function gives the power spectrum; knowledge of one is equivalent to the other. An important example of this is white noise: a memoryless process with a delta function autocorrelation will have a flat power spectrum, regardless of the probability distribution

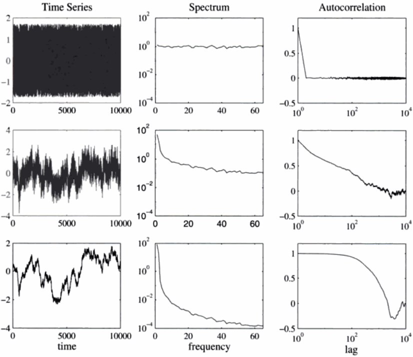


Figure 3.1. Illustration of the Wiener–Khinchin Theorem: as the power spectrum decays more quickly, the autocorrelation function decays more slowly.

for the signal. As the autocorrelation function decays more slowly, the power spectrum will decay more quickly (Figure 3.1).

Taking $\tau = 0$ in the Wiener–Khinchin Theorem yields *Parseval's Theorem*:

$$\langle x(t)x(t-\tau)\rangle = \int_{-\infty}^{\infty} S(f)e^{-i2\pi f\tau}\, df = \int_{-\infty}^{\infty} \langle |X(f)|^2\rangle e^{-i2\pi f\tau}\, df$$

$$\Rightarrow \langle |x|^2(t)\rangle = \int_{-\infty}^{\infty} \langle |X(f)|^2\rangle\, df \quad . \tag{3.12}$$

The average value of the square of the signal (which is equal to the variance if the signal has zero mean) is equal to the integral of the power spectral density. This means that true white noise has an infinite variance in the time domain, although the finite bandwidth of any real system will roll off the frequency response, and hence determine the variance of the measured signal. If the division by T is left off in the limiting process defining the averages on both sides of Parseval's Theorem, then it reads that the total energy in the signal equals the total energy in the spectrum (the integral of the square of the magnitude).

3.2 PROBABILITY DISTRIBUTIONS

So far we have taken the probability distribution $p(x)$ to be arbitrary. In practice, three probability distributions recur so frequently that they receive most attention: *binomial*, *Poisson* and *Gaussian*. Their popularity is due in equal parts to the common conditions that give rise to them and to the convenience of working with them. The latter reason sometimes outweighs the former, leading these distributions to be used far from where they apply. For example, many physical system have *long-tailed distributions* that fall off much more slowly than these ones do [Crisanti *et al.*, 1993; Boguna & Corral, 1997].

3.2.1 Binomial

Consider many trials of an event that can have one outcome with probability p (such as flipping a coin and seeing a head), and an alternative with probability $1-p$ (such as seeing a tail). In n trials, the probability $p_n(x)$ to see x heads and $n-x$ tails, independent of the particular order in which they were seen, is found by adding up the probability for each outcome times the number of equivalent arrangements:

$$p_n(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad , \tag{3.13}$$

where

$$\binom{n}{x} = \frac{n!}{(n-x)!\, x!} \tag{3.14}$$

(read "n choose x"). This is the *binomial distribution*. The second line follows by dividing the total number of distinct arrangements of n objects ($n!$) by the number of equivalent distinct arrangements of heads $x!$ and tails $(n-x)!$. The easiest way to convince yourself that this is correct is to exhaustively count the possibilites for a small case.

3.2.2 Poisson

Now consider events such as radioactive decays that occur randomly in time. Divide time into n very small intervals so that there are either no decays or one decay in any one interval, and let p be the probability of seeing a decay in an interval. If the total number of events that occur in a given time is recorded, and this is repeated many times to form an ensemble of measurements, then the distribution of the total number of events recorded will be given by the binomial distribution. If the number of intervals n is large, and the probability p is small, the binomial distribution can be approximated by using $\ln(1+x) \approx x$ for small x and *Stirling's approximation* for large n:

$$\begin{aligned} n! &\approx \sqrt{2\pi}\, n^{n+\frac{1}{2}} e^{-n} \\ \ln n! &\approx n \ln n - n \quad , \end{aligned} \tag{3.15}$$

to find the *Poisson distribution* (Problem 3.1):

$$p(x) = \frac{e^{-N} N^x}{x!} \quad , \tag{3.16}$$

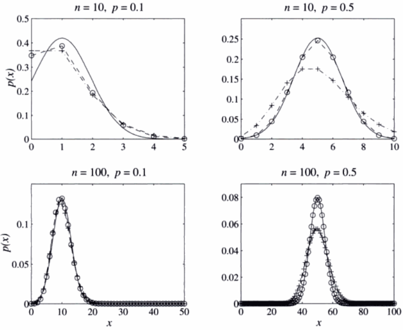


Figure 3.2. Comparison of the binomial (○), Poisson (+) and Gaussian (−) distributions: n is the number of trials, and p is the probability of seeing an event. By definition, the binomial distribution is correct. For a small probability of seeing an event, the Poisson distribution is a better approximation (although the difference is small for a large number of events), while for a large probability of seeing an event the Gaussian distribution is closer.

any distribution. For these reasons, it is often safe (and certainly common) to assume that an unknown distribution is Gaussian.

The Fourier transform of a Gaussian has a particularly simple form, namely a Gaussian with the inverse of the variance

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-x^2/2\sigma^2} e^{ikx}\, dx = e^{-k^2\sigma^2/2} \quad . \tag{3.23}$$

Remember this: you should never need to look up the transform of a Gaussian, just invert the variance. Because of this relationship, the product of the variance of a Gaussian and the variance of its Fourier transform will be a constant; this is the origin of many classical and quantum uncertainty relationships.

Figure 3.2 compares the binomial, Poisson, and Gaussian distributions for $n = 10$ and 100, and for $p = 0.1$ and 0.5, showing where they are and are not good approximations.

will become vanishing small compared to the lower-order terms in the limit $N \to \infty$. The last line follows because an exponential can be written as

$$\lim_{N\to\infty}\left(1+\frac{x}{N}\right)^N = e^x \quad , \tag{3.28}$$

which can be verified by comparing the Taylor series of both sides. To find the probability distribution for y we now take the inverse transform

$$\begin{aligned} p(y-\langle x\rangle) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-k^2\sigma^2/2N} e^{-ik(y-\langle x\rangle)}\, dk \\ &= \sqrt{\frac{N}{2\pi\sigma^2}} e^{-N(y-\langle x\rangle)^2/2\sigma^2} \end{aligned} \tag{3.29}$$

(remember that the Fourier transform of a Gaussian is also a Gaussian). This proves the Central Limit Theorem [Feller, 1974]. The average of N iid variables has a Gaussian distribution, with a standard deviation $\sigma/\sqrt{N}$ reduced by the square root of the number of variables just as with Poisson statistics. It can be a surprisingly good approximation even with just tens of samples. The Central Limit Theorem also contains the *Law of Large Numbers*: in the limit $N \to \infty$, the average of N random variables approaches the mean of their distribution. Although this might appear to be a trivial insight, lurking behind it is the compressibility of data that is so important to digital coding (Section 4.1).

3.3 NOISE MECHANISMS

Now that we've seen something about how to describe random systems we will turn to a quantitative discussion of some of the most important fundamental noise mechanisms: *shot noise*, *Johnson noise*, and *1/f noise*. Chapter 13 will consider other practical sources of noise, such as interference from unwanted signals.

3.3.1 Shot Noise

A current, such as electrons in a wire or rain on a roof, is made up of the discrete arrival of many carriers. If their interactions can be ignored so that they arrive independently, this is an example of a Poisson process. For an electrical signal, the average current is $\langle I\rangle = qN/T$ for N electrons with charge q arriving in a time T. If the electrons arrive far enough apart so that the duration during which they arrive is small compared to the time between the arrival of successive electrons, then the current can be approximated as a sum of delta functions.

$$I(t) = q\sum_{n=1}^{N}\delta(t-t_n) \quad , \tag{3.30}$$

where t_n is the arrival time for the nth electron. The Fourier transform of this impulse train is.

$$\begin{aligned} I(f) &= \lim_{T\to\infty}\int_{-T/2}^{T/2} e^{i2\pi ft}\, q\sum_{n=1}^{N}\delta(t-t_n)\, dt \\ &= q\sum_{n=1}^{N} e^{i2\pi ft_n} \quad . \end{aligned} \tag{3.31}$$

Therefore, the power spectrum is

$$
\begin{aligned}
S_I(f) &= \langle I(f)I^*(f)\rangle \\
&= \lim_{T\to\infty} \frac{q^2}{T}\left(\sum_{n=1}^{N} e^{i2\pi f t_n} \sum_{m=1}^{N} e^{-i2\pi f t_m}\right) \\
&= \lim_{T\to\infty} \frac{q^2 N}{T} \\
&= q\langle I\rangle
\end{aligned}
\tag{3.32}
$$

(the cross terms $n \neq m$ vanish in the expectation because their times are independent). We see that the power spectrum of carrier arrivals is white (flat) and that the magnitude is linearly proportional to the current. This is called *shot noise* or *Schottky noise*. If the carriers do not really arrive as delta functions then the broadening of the impulses will roll the spectrum off for high frequencies, so the flat power spectrum is a good approximation up to the inverse of the characteristic times in the system.

To find the fluctuations associated with shot noise, we can use Parseval's Theorem to relate the average total energy in the spectrum to the average variance. If the bandwidth of the system is infinite this variance will be infinite, because for ideal shot noise there is equal power at all frequencies. Any real measurement system will have a finite bandwidth, and this determines the amplitude of the noise. Multiplying the power spectrum by $2\Delta f$, where Δf is the bandwidth in hertz and the factor of 2 comes from including both positive and negative frequencies,

$$
\langle I^2_{\text{noise}}\rangle = 2q\langle I\rangle\Delta f \quad . \tag{3.33}
$$

Shot noise will be important only if the number of carriers is small enough for the rate of arrival to be discernible; Problem 3.2 looks at this limit for detecting light.

3.3.2 Johnson Noise

Johnson (or *Nyquist*) noise is the noise associated with the relaxation of thermal fluctuations in a resistor. Small voltage fluctuations are caused by the thermal motion of the electrons, which then relax back through the resistance. We will calculate this in Section 3.4.3, but the result is simple:

$$
\langle V^2_{\text{noise}}\rangle = 4kTR\Delta f \tag{3.34}
$$

(where R is resistance, Δf is the bandwidth of the measuring system, T is the temperature, and k is Boltzmann's constant). Once again, this is white noise, but unlike shot noise it is independent of the current. The resistor is acting almost like a battery, driven by thermodynamic fluctuations. The voltage produced by these fluctuations is very real and very important: it sets a basic limit on the performance of many kinds of electronics. Unfortunately, it is not possible to take advantage of Johnson noise by rectifying the fluctuating voltage across a diode to use a resistor as a power source (hint: what temperature is the diode?).

Johnson noise is an example of a *fluctuation–dissipation* relationship (Section 3.4.3) – the size of a system's thermodynamic fluctations is closely related to the rate at which the system relaxes to equilibrium from a perturbation. A system that is more strongly damped has smaller fluctuations, but it dissipates more energy.

3.3.3 $1/f$ Noise and Switching Noise

In a wide range of transport processes, from electrons in resistors, to cars on the highway, to notes in music, the power spectrum diverges at low frequencies inversely proportionally to frequency: $S(f) \propto f^{-1}$. Because such *1/f noise* is scale-invariant (the spectrum looks the same at all time scales [Mandelbrot, 1983]) and is so ubiquitous, many people have been lured to search for profound general explanations for the many particular examples. While this has led to some rather bizarre ideas, there is a reasonable theory for the important case of electrical $1/f$ noise.

In a conductor there are usually many types of defects, such as lattice vacancies or dopant atoms. Typically, the defects can be in a few different inequivalent types of sites in the material, which have different energies. This means that there is a probability for a defect to be thermally excited into a higher-energy state, and then relax down to the lower-energy state. Because the different sites can have different scattering cross-sections for the electron current, this results in a fluctuation in the conductivity of the material. A process that is thermally activated between two states, with a characteristic time τ to relax from the excited state, has a Lorentzian power spectrum of the form

$$S(f) = \frac{2\tau}{1 + (2\pi f \tau)^2} \tag{3.35}$$

(we will derive this in Problem 3.4). If there is a distribution of activation times $p(\tau)$ instead of a single activation time in the material, and if the activated scatterers don't interact with each other, then the spectrum will be an integral over this:

$$S(f) = \int_0^\infty \frac{2\tau}{1 + (2\pi f \tau)^2}\, p(\tau)\, d\tau \quad . \tag{3.36}$$

If the probability of the defect having an energy equal to a barrier height E goes as $e^{-E/kT}$ (Section 3.4), then the characteristic time τ to be excited over the barrier will be inversely proportional to probability

$$\tau = \tau_0 e^{E/kT} \quad . \tag{3.37}$$

This is called a *thermally activated* process. If the distribution of barrier heights $p(E)$ is flat then $p(\tau) \propto 1/\tau$, and putting this into equation (3.36) shows that $S(f) \propto 1/f$ (Problem 3.4) [Dutta & Horn, 1981].

This is the origin of $1/f$ noise: scatterers with a roughly flat distribution of activation energies. Cooling a sample to a low enough temperature can turn off the higher-energy scatterers and reveal the individual Lorentzian components in the spectrum [Rogers & Buhrman, 1984]. In this regime, the noise signal in time is made up of jumps between discrete values, called *switching noise*. This can be seen unexpectedly and intermittently at room temperature, for example if a device has a very bad wire-bond so that the current passes through a narrow constriction.

Unlike Johnson noise, $1/f$ noise is proportional to the current in the material because it is a conductivity rather than a voltage fluctuation, and it increases as the cross-sectional

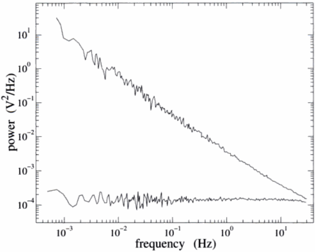


Figure 3.3. Noise in a 50 Ω resistor with and without a current.

area of the material is decreased because the relative influence of a single defect is greater. That is why $1/f$ noise is greater in carbon resistors, which have many small contacts between grains, than in metal film resistors. Low-noise switches have large contact areas, and wiping connections that slide against each other as the switch is closed, to make sure that the conduction is not constrained to small channels.

The power spectrum of the noise from a resistor will be flat because of Johnson noise if there is no current flowing; as the current is increased the $1/f$ noise will appear, and the frequency below which it is larger than the Johnson noise will depend on the applied current as well as on the details of the material. $1/f$ noise is not an intrinsic property: the magnitude is a function of how a particular sample is prepared. Figure 3.3 shows the Johnson and $1/f$ noise for a carbon resistor. Because $1/f$ noise diverges at low frequencies, it sets a time limit below which measurements cannot be made; a common technique to avoid $1/f$ noise is to modulate the signal up to a higher frequency (we will discuss this in Chapter 13).

3.3.4 Amplifier Noise

Any device that detects a signal must contend with these noise mechanisms in its workings. Johnson noise leads to the generation of voltage noise by an amplifier. Since the power spectral density is flat, the mean square noise magnitude will be proportional to the bandwidth, or the *Root Mean Square* (*RMS*) magnitude will increase as the square root of the bandwidth. The latter quantity is what is conventionally used to characterize an amplifier; for a low-noise device it can be on the order of 1 nV/$\sqrt{\text{Hz}}$. Likewise, shot noise is responsible for the generaton of current noise at an amplifier's output; this is also flat and for a low-noise amplifier can be on the order of 1 pA/$\sqrt{\text{Hz}}$.

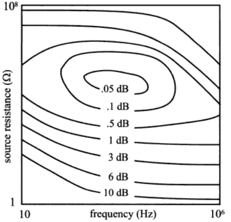


Figure 3.4. Noise contours for a low-noise amplifier.

large as the device. Since inelastic scattering is the origin of resistance and hence of the thermodynamic coupling of the conduction electrons to the material, this means that the noise temperature can be much lower than room temperature. In the best devices it gets down to just a few kelvins. One of the places where this sensitivity is particularly important is for detecting the weak signals from space for satellite communications and radio astronomy.

3.4 THERMODYNAMICS AND NOISE

Thermal fluctuations and noise are intimately related. This section turns to a more general discussion of this connection, starting with a brief review of macroscopic thermodynamics and its origin in microscopic statistical mechanics, and then looking at the Equipartition Theorem (which relates temperature to the average energy stored in a system's degrees of freedom) and the Fluctuation–Dissipation Theorem (which relates fluctuations to the dissipation in a system).

3.4.1 Thermodynamics and Statistical Mechanics

A thermodynamic system can be described by a *temperature* T, an *internal energy* E, and an *entropy* S. The internal energy is the sum off all of the energy stored in all of the degrees of freedom of the system. The entropy provides a relationship between heat and temperature: if the system is kept at a constant temperature, and a heat current δQ flows into or out of the system, the change in entropy is

$$\delta Q = T\,dS \quad . \tag{3.41}$$

This is written as δQ rather than dQ because energy that flows in and increases the entropy of a system cannot be reversibly recovered to do work. In any spontaneous the constraints that we impose. Justifying this essentially experimental fact is the subject of endless mathematical if not mystical discussion; Boltzmann's *H-Theorem* provides a derivation in the context of scattering in a dilute gas [Reichl, 1998].

For the canonical ensemble there are two constraints: the probability distribution must be normalized

$$\sum_{i=1}^{\Omega} p_i = 1 \quad , \tag{3.47}$$

and the average energy must be a constant E

$$\sum_{i=1}^{\Omega} E_i p_i = E \quad . \tag{3.48}$$

To do a constrained maximization we will use the method of *Lagrange multipliers*. Define a quantity I to be the entropy plus Lagrange multipliers times the constraint equations

$$I = -k \sum_{i=1}^{\Omega} p_i \log p_i + \lambda_1 \sum_{i=1}^{\Omega} p_i + \lambda_2 \sum_{i=1}^{\Omega} E_i p_i \quad . \tag{3.49}$$

We want to find the values for the p_i's that make this extremal:

$$\frac{\partial I}{\partial p_i} = 0 \quad . \tag{3.50}$$

We can do this because the two terms that we've added are just constants, Equations (3.47) and (3.48); we just need to choose the values of the Lagrange multipliers to make sure that they have the right values. Solving,

$$\frac{\partial I}{\partial p_i} = 0 = -k \log p_i - k + \lambda_1 + \lambda_2 E_i \tag{3.51}$$

$$\Rightarrow p_i = e^{(\lambda_1/k)+(\lambda_2 E_i/k)-1} \quad . \tag{3.52}$$

If we sum this over i,

$$\sum_{i=1}^{\Omega} p_i = 1 = e^{\lambda_1/k-1} \sum_{i=1}^{\Omega} e^{\lambda_2 E_i/k} \quad . \tag{3.53}$$

This can be rearranged to define the *partition function* $\mathcal{Z}$

$$\mathcal{Z} \equiv e^{1-\lambda_1/k} = \sum_{i=1}^{\Omega} e^{\lambda_2 E_i/k} \quad . \tag{3.54}$$

Another equation follows from multiplying equation (3.51) by p_i and summing:

$$\sum_{i=1}^{\Omega} p_i \frac{\partial I}{\partial p_i} = S - k + \lambda_1 + \lambda_2 E = 0 \quad . \tag{3.55}$$

Since

$$\mathcal{Z} = e^{1-\lambda_1/k} \quad , \tag{3.56}$$

$$k \log \mathcal{Z} = k - \lambda_1 \quad , \tag{3.57}$$

and so equation (3.55) can be written as

$$S - k \log \mathcal{Z} + \lambda_2 E = 0 \quad . \tag{3.58}$$

Comparing this to the definition of the free energy $A = E - TS$, we see that

$$S - \underbrace{k \log \mathcal{Z}}_{-A/T} + \underbrace{\lambda_2}_{-1/T} E = 0 \quad . \tag{3.59}$$

This provides a connection between the macroscopic thermodynamic quantities and the microscopic statistical mechanical ones.

Putting the value of λ_2 into equation (3.54) shows that the partition function is given by

$$\mathcal{Z} = \sum_{i=1}^{\Omega} e^{-E_i/kT} \equiv \sum_{i=1}^{\Omega} e^{-\beta E_i} \quad . \tag{3.60}$$

Returning to equation (3.52) we see that

$$p_i = e^{\lambda_1/k-1} e^{-E_i/kT} = \frac{e^{-E_i/kT}}{\mathcal{Z}} \quad . \tag{3.61}$$

In terms of this, the expected value of a function f_i that depends on the state of the system is

$$\langle f \rangle = \sum_{i=1}^{\Omega} f_i p_i = \frac{\sum_{i=1}^{\Omega} f_i e^{-E_i/kT}}{\mathcal{Z}} \tag{3.62}$$

3.4.2 Equipartition Theorem

The *Equipartition Theorem* is a simple, broadly applicable result that can give the magnitude of the thermal fluctuations associated with energy storage in independent degrees of freedom of a system. Assume that the state of a system is specified by variables $x_0, \ldots, x_n$, and that the internal energy of the system is given in terms of them by

$$E = E(x_0, \ldots, x_n) \quad . \tag{3.63}$$

Now consider the case where one of the degrees of freedom splits off additively in the energy:

$$E = E_0(x_0) + E_1(x_1, \ldots, x_n) \quad . \tag{3.64}$$

E might be the energy in a circuit, and $E_0 = CV_0^2/2$ the energy in a particular capacitor in terms of the voltage V_0 across it, or $E_0 = mv_0^2/2$ the kinetic energy of one particle in terms of its velocity v_0.

If we now assume that the overall system is in equilibrium at a temperature T, the expectation value for E_0 is given by the canonical statistical mechanical distribution (here taken as an integral instead of a discrete sum for a continuous system)

$$
\begin{aligned}
\langle E_0 \rangle &= \frac{\int_{-\infty}^{\infty} e^{-\beta E(x_0,\ldots,x_n)} E_0(x_0)\, dx_0 \cdots dx_n}{\int_{-\infty}^{\infty} e^{-\beta E(x_0,\ldots,x_n)}\, dx_0 \cdots dx_n} \qquad (\beta \equiv kT) \\
&= \frac{\int_{-\infty}^{\infty} e^{-\beta[E_0(x_0)+E_1(x_1,\ldots,x_n)]} E_0(x_0)\, dx_0 \cdots dx_n}{\int_{-\infty}^{\infty} e^{-\beta[E_0(x_0)+E_1(x_1,\ldots,x_n)]}\, dx_0 \cdots dx_n} \\
&= \frac{\int_{-\infty}^{\infty} e^{-\beta E_0(x_0)} E_0(x_0)\, dx_0 \int_{-\infty}^{\infty} e^{-\beta E_1(x_1,\ldots,x_n)}\, dx_1 \cdots dx_n}{\int_{-\infty}^{\infty} e^{-\beta E_0(x_0)}\, dx_0 \int_{-\infty}^{\infty} e^{-\beta E_1(x_1,\ldots,x_n)}\, dx_1 \cdots dx_n} \\
&= \frac{\int_{-\infty}^{\infty} e^{-\beta E_0(x_0)} E_0(x_0)\, dx_0}{\int_{-\infty}^{\infty} e^{-\beta E_0(x_0)}\, dx_0} \\
&= -\frac{\partial}{\partial \beta} \ln \int_{-\infty}^{\infty} e^{-\beta E_0(x_0)}\, dx_0 \quad . \qquad (3.65)
\end{aligned}
$$

If $E_0 = a x_0^2$ for some constant a, we can simplify the integral further:

$$
\begin{aligned}
\langle E_0 \rangle &= -\frac{\partial}{\partial \beta} \ln \int_{-\infty}^{\infty} e^{-\beta E_0(x_0)}\, dx_0 \\
&= -\frac{\partial}{\partial \beta} \ln \int_{-\infty}^{\infty} e^{-\beta a x_0^2}\, dx_0 \\
&= -\frac{\partial}{\partial \beta} \ln \left[\frac{1}{\sqrt{\beta}} \int_{-\infty}^{\infty} e^{-a y^2}\, dy \right] \qquad (y^2 \equiv \beta x_0^2) \\
&= -\frac{\partial}{\partial \beta} \left[-\frac{1}{2} \ln \beta + \ln \int_{-\infty}^{\infty} e^{-a y^2}\, dy \right] \\
a\langle x_0^2 \rangle &= \frac{1}{2} kT \quad . \qquad (3.66)
\end{aligned}
$$

Each independent thermalized quadratic degree of freedom has an average energy of $kT/2$ due to fluctuations.

3.4.3 Fluctuation–Dissipation Theorem

The Equipartition Theorem relates the size of thermal fluctuations to the energy stored in independent degrees of freedom of a system; the Fluctuation–Dissipation Theorem relates the thermal fluctuations to the amount of dissipation. We will start with a simple example and then discuss the more general theory. Consider an ideal inductor L connected in parallel with a resistor R. Because of thermal fluctuations there will be a voltage across the resistor; model that by a fluctuating voltage source V in series with a noiseless resistor (Figure 3.5).

In Chapter 6 we will show that the energy stored in an inductor is $LI^2/2$. Since the inductor is the only energy storage element, from the equipartition theorem we know what the current across it due to thermal fluctuations must be:

$$
\left\langle \frac{1}{2} L I^2 \right\rangle = \frac{1}{2} kT \quad . \qquad (3.67)
$$

Ohm's Law (Section 6.1.3) still applies, so this current must also be equal to the fluctuating thermal voltage divided by the total impedance Z of the circuit. Written in terms

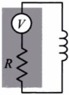


Figure 3.5. Resistor modeled as a fluctuating voltage source in series with a noiseless resistor, connected in parallel with an inductor.

of the frequency components,

$$I(\omega) = \frac{V(\omega)}{Z(\omega)} = \frac{V(\omega)}{R + i\omega L(\omega)} \tag{3.68}$$

(we will explain why the impedance of an inductor is $i\omega L$ when we derive the circuit equations from Maxwell's equations). Writing the equipartition result in terms of frequency components,

$$\begin{aligned}
\frac{1}{2}kT &= \left\langle \frac{1}{2}LI^2 \right\rangle = \frac{L}{2}\langle I^2 \rangle \\
&= \frac{L}{2}\int_{-\infty}^{\infty} \left\langle |I(\omega)|^2 \right\rangle \, d\omega \quad \text{(Parseval's Theorem)} \\
&= \frac{L}{2}\int_{-\infty}^{\infty} \left\langle \frac{|V(\omega)|^2}{|Z(\omega)|^2} \right\rangle \, d\omega \\
&= \frac{L}{2}\int_{-\infty}^{\infty} \frac{\left\langle |V(\omega)|^2 \right\rangle}{R^2 + \omega^2 L^2} \, d\omega \quad .
\end{aligned} \tag{3.69}$$

Since this is assumed to be an ideal resistor with no time constant from an inductive or capacitative component, it's a reasonable assumption to take the fluctuating voltage V to have a delta function autocorrelation (this can be justified by a microscopic derivation). And since that implies that the power spectrum of the fluctuations is flat, V does not depend on ω and can come out of the intergral:

$$\frac{1}{2}kT = \frac{L\langle V^2(\omega)\rangle}{2}\int_{-\infty}^{\infty} \frac{1}{R^2 + \omega^2 L^2} \, d\omega \quad . \tag{3.70}$$

This integration can then be done analytically,

$$\frac{1}{2}kT = \frac{\pi\langle V^2(\omega)\rangle}{2R} \quad . \tag{3.71}$$

Therefore,

$$\begin{aligned}
\frac{\pi\langle V^2(\omega)\rangle}{2R} &= \frac{1}{2}kT \\
\langle V^2(\omega)\rangle &= \frac{kTR}{\pi} \\
\langle V^2(f)\rangle &= 4kTR \quad .
\end{aligned} \tag{3.72}$$

In the last line there is a factor of 2π to convert from radian per second to cycles per gives energy, and that energy per time gives power. Therefore multiplying the driving force dS/dx by dx and dividing by dt gives the power P being dissipated,

$$P = \frac{dS}{dx}\frac{dx}{dt} = R\left(\frac{dx}{dt}\right)^2 \quad . \tag{3.79}$$

Therefore equation (3.77) shows that

$$P = k^2\frac{\alpha^2}{R}\langle x^2\rangle = k^2\frac{\alpha}{R} \quad . \tag{3.80}$$

If the entropy is sharply peaked (α large relative to R), then the fluctuations will be small but the dissipation will be large. If the entropy is flatter (α small), the fluctuations will be large but the dissipation will be small. A related equation is found by multiplying both sides of equation (3.77) by x and averaging:

$$R\;\underbrace{x\frac{dx}{dt}}_{\frac{1}{2}\frac{dx^2}{dt}} = -k\alpha x^2$$

$$\frac{d\langle x^2\rangle}{dt} = -2k\frac{\alpha}{R}\langle x^2\rangle \quad . \tag{3.81}$$

If the system is perturbed, the variance also relaxes at a rate proportional to α/R. It doesn't go to zero, of course, because we've left off the noise source term in the Langevin equation that drives the fluctuations.

Equation (3.80) is a simple example of the *Fluctuation–Dissipation Theorem*. The generalization is straightforward to systems with more degrees of freedom [Montroll & Lebowitz, 1987; Reichl, 1998] and to quantum systems [Balian, 1991]. In higher dimensions the relaxation constant R becomes a matrix, and if the system has time reversal invariance so that the governing equations are the same if $t \rightarrow -t$ then this matrix is symmetrical ($R_{ij} = R_{ji}$, called the *Onsager reciprocal relationship*).

The fluctuation dissipation theorem can be understood by remembering that a change in entropy is associated with a heat current $\delta Q = TdS$; if the entropy is sharply peaked then the fluctuations lead to larger changes in the entropy. This is an essential tradeoff in the design of any system: the faster and more accurately you want it to do something, the more power it will require. For example, one of the most important lessons in the design of low-power electronics is to make sure that the system does not produce results any faster than they are needed. This also shows why, without knowing anything else about electronics, low-noise amplifiers require more power than noisy ones.

3.5 SELECTED REFERENCES

[Feller, 1968] Feller, William. (1968). *An Introduction to Probability Theory and Its Applications*. 3rd edn. New York: Wiley.

[Feller, 1974] Feller, William. (1974). *An Introduction to Probability Theory and Its Applications*. 2nd edn. Vol. II. New York: Wiley.

A definitive probability reference.

4 Information in Physical Systems

What is information? A good answer is that information is what you don't already know. You do not learn much from being told that the sun will rise tomorrow morning; you learn a great deal if you are told that it will not. Information theory quantifies this intuitive notion of surprise. Its primary success is an explanation of how noise and energy limit the amount of information that can be represented in a physical system, which in turn provides insight into how to efficiently manipulate information in the system.

In the last chapter we met some of the many ways that devices can introduce noise into a signal, effectively adding unwanted information to it. This process can be abstracted into the concept of a *communications channel* that accepts an input and then generates an output. A telephone connection is a channel, as is the writing and subsequent reading of bits on a hard disk. In all cases there is assumed to be a set of known input symbols (such as 0 and 1), possibly a device that maps them into other symbols in order to satisfy constraints of the channel, the channel itself which has some probability for modifying the message due to noise or other errors, and possibly a decoder that turns the received symbols into an output set. We will assume that the types of messages and types of channel errors are sufficiently stationary to be able to define probability distributions $p(x)$ to see an input message x, and $p(y|x)$ for the channel to deliver a y if it is given an input x. This also assumes that the channel has no memory so that the probability distribution depends only on the current message. These are important assumptions: the results of this chapter will not apply to non-stationary systems.

4.1 INFORMATION

Let x be a random variable that takes on X possible values indexed by $i = 1, \ldots, X$, and let the probability of seeing the ith value be p_i. For example, x could be the letters of the alphabet, and p_i could be the probability to see letter i. How much information is there on average in a value of x drawn from this distribution? If there is only one possible value for x then we learn very little from successive observations because we already know everything; if all values are equally likely we learn as much as possible from each observation because we start out knowing nothing. An information functional $H(p)$ (a functional is a function of a function) that captures this intuitive notion should have the following reasonable properties:

- $H(p)$ is continuous in p. Small changes in the distribution should lead to small changes in the information.
- $H(p) \geq 0$, and $H(p) = 0$ if and only if just one p_i is non-zero. You always learn something unless you already know everything.
- $H(p) \leq C(X)$, where $C(X)$ is a constant that depends on the number of possible values X, with $H(p) = C(X)$ when all values are equally likely, and $X' > X \Rightarrow C(X') > C(X)$. The more options there are, the less you know about what will happen next.
- If x is drawn from a distribution p and y is independently drawn from a distribution q, then $H(p, q) = H(p) + H(q)$, where $H(p, q)$ is the information associated with seeing a pair (x, y). The information in independent events is the sum of the information in the events individually.

While it might appear that this list is not sufficient to define $H(p)$, it can be shown [Ash, 1990] that these desired properties are uniquely satisfied by the function

$$H(p) = -\sum_{i=1}^{X} p_i \log p_i \quad . \tag{4.1}$$

This is the definition of the *entropy* of a probability distribution, the same definition that was used in the last chapter in statistical mechanics. To make the dependence on x clear, we will usually write this as $H(x)$ instead of $H(p(x))$ or $H(p)$. The choice of the base of the logarithm is arbitrary; if the base is 2 then the entropy is measured in *bits*, and if it is base e then the entropy units are called *nats* for the *natural logarithm*. Note that to change an entropy formula from bits to nats you just change the logarithms from $\log_2$ to $\log_e$, and so unless otherwise noted the base of the logarithms in this chapter is arbitrary.

Now consider a string of N samples $(x_1, \ldots, x_N)$ drawn from p, and let N_i be the number of times that the ith value of x was actually seen. Because of the independence of the observations, the probability to see a particular string is the product of the individual probabilities

$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} p(x_n) \quad . \tag{4.2}$$

This product of terms can be regrouped in terms of the possible values of x,

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{X} p_i^{N_i} \quad . \tag{4.3}$$

Taking the log and multiplying both sides by $-1/N$ then lets this be rewritten as

$$\begin{aligned} -\frac{1}{N} \log p(x_1, \ldots, x_N) &= -\frac{1}{N} \log \prod_{i=1}^{X} p_i^{N_i} \\ &= -\sum_{i=1}^{X} \frac{N_i}{N} \log p_i \\ &\approx -\sum_{i=1}^{X} p_i \log p_i \\ &= H(x) \quad . \end{aligned} \tag{4.4}$$

The third line follows from the Law of Large Numbers (Section 3.2.4): as $N \to \infty$, $N_i/N \to p_i$. Equation (4.4) can be inverted to show that

$$p(x_1, \ldots, x_N) \approx 2^{-NH(x)} \tag{4.5}$$

(taking the entropy to be defined base 2). Something remarkable has happened: the probability of seeing a particular long string is independent of the elements of that string. This is called the *Asymptotic Equipartition Property* (*AEP*). Since the probability of occurrence for a string is a constant, its inverse $1/p = 2^{NH(x)}$ gives the effective number of strings of that length. However, the actual number of strings is larger, equal to

$$X^N = 2^{N \log_2 X} \quad . \tag{4.6}$$

The difference between these two values is what makes data compression possible. It has two very important implications [Blahut, 1988]:

- Since samples drawn from the distribution can on average be described by $H(x)$ bits rather than $\log_2 X$ bits, a coder can exploit the difference to store or transmit the string with $NH(x)$ bits. This is *Shannon's First Coding Theorem*, also called the *Source Coding Theorem* or the *Noiseless Coding Theorem*.
- The compressibility of a typical string is made possible by the vanishing probability to see rare strings, the ones that violate the Law of Large Numbers. In the unlikely event that such a string appears the coding will fail and a longer representation must be used. Because the Law of Large Numbers provides an increasingly tight bound on this occurrence as the number of samples increases, the failure probability can be made arbitrarily small by using a long enough string. This is the *Shannon–McMillan Theorem*.

Because the entropy is a maximum for a flat distribution, an efficient coder will represent information with this distribution. This is why modems "hiss": they make best use of the telephone channel if the information being sent appears to be as random as possible. The value of randomness in improving a system's performance will recur throughout this book, particularly in Chapter 13.

We see that the entropy (base 2) gives the average number of bits that are required to describe a sample drawn from the distribution. Since the entropy is equal to

$$-\sum_{i=1}^{X} p_i \log p_i = \langle -\log p_i \rangle \tag{4.7}$$

it is natural to interpret $-\log p_i$ as the information in seeing event p_i, and the entropy as the expected value of that information.

Entropy can be applied to systems with more degrees of freedom. The joint entropy for two variables with a joint distribution $p(x, y)$ is

$$H(x, y) = -\sum_x \sum_y p(x, y) \log p(x, y) \quad . \tag{4.8}$$

This can be rewritten as

$$
\begin{aligned}
H(x,y) &= -\sum_x \sum_y p(x,y) \log p(x,y) \\
&= -\sum_x \sum_y p(x,y) \log[p(x|y)p(y)] \\
&= -\sum_x \sum_y p(x,y) \log p(x|y) - \sum_x \sum_y p(x,y) \log p(y) \\
&= -\sum_x \sum_y p(x,y) \log p(x|y) - \sum_y p(y) \log p(y) \\
&= H(x|y) + H(y)
\end{aligned}
\tag{4.9}
$$

by using *Bayes' rule* $p(x,y) = p(x|y)p(y)$. The entropy in a conditional distribution $H(x|y)$ is the expected value of the information $\langle -\log p(x|y) \rangle$. The entropy of both variables equals the entropy of one of them plus the entropy of the other one given the observation of the first.

The *mutual information* between two variables is defined to be the information in them taken separately minus the information in them taken together

$$
\begin{aligned}
I(x,y) &= H(x) + H(y) - H(x,y) \\
&= H(y) - H(y|x) \\
&= H(x) - H(x|y) \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\end{aligned}
\tag{4.10}
$$

(these different forms are shown to be equal in Problem 4.2). This measures how many bits on average one sample tells you about the other. It vanishes if the variables are independent, and it is equal to the information in one of them if they are completely dependent. The mutual information can be viewed as an information-theoretic analog of the cross-correlation function $\langle x(t)y(t) \rangle$, but the latter is useful only for measuring the overlap among signals from linear systems [Gershenfeld, 1993].

In a sequence of N values $(x_1, x_2, \ldots, x_N)$ the *joint* (or *block entropy*)

$$
H_N(x) = -\sum_{x_1} \sum_{x_2} \cdots \sum_{x_N} p(x_1, x_2, \ldots, x_N) \log p(x_1, x_2, \ldots, x_N) \tag{4.11}
$$

is the average number of bits needed to describe the string. The limiting rate at which this grows

$$
h(x) = \lim_{N\to\infty} \frac{1}{N} H_N(x) = \lim_{N\to\infty} H_{N+1} - H_N \tag{4.12}
$$

is called the *source entropy*. It is the rate at which the system generates new information.

So far we've been discussing random variables that can take on a discrete set of values; defining entropy for continuous variables requires some care. If x is a real number, then $p(x)\,dx$ is the probability to see a value between x and $x + dx$. The information in such an observation is given by its logarithm, $-\log[p(x)\,dx] = -\log p(x) - \log dx$. As $dx \to 0$ this will diverge! The divergence is in fact the correct answer, because a single real number can contain an infinite amount of information if it can be specified to any

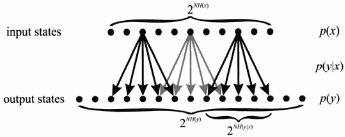


Figure 4.1. Effective number of states input to, added by, and output from a channel.

that are input to a channel specified by $p(y|x)$. On average each sample contains $H(x)$ bits of information, so this input string of N symbols can represent roughly $2^{NH(x)}$ different states. After being sent through the channel an output string $(y_1, y_2, \ldots, y_N)$ can represent $2^{NH(y)}$ states. However, it is possible that because of noise in the channel different input states can produce the same output state and hence garble the message; $2^{NH(y|x)}$ is the average number of different output states that are produced by an input state, the extra information in y given knowledge of x. In order to make sure that each input state typically leads to only one output state it is necessary to reduce the number of allowable output states by the excess information generated by the channel (Figure 4.1)

$$\frac{2^{NH(y)}}{2^{NH(y|x)}} = 2^{N[H(y)-H(y|x)]} = 2^{NI(x,y)} \quad . \tag{4.17}$$

We see that the probability distribution that maximizes the mutual information between the input and the output leads to the maximum number of distinct messages that can reliably be sent through the channel. The *channel capacity* is this maximum bit rate:

$$C = \max_{p(x)} I(x, y) \quad . \tag{4.18}$$

Applying the Shannon–McMillan Theorem to the input and output of the channel taken together shows that, if the data rate is below the channel capacity and the block length is long enough, then messages can be decoded with an arbitrarily small error. On the other hand, it is impossible to send data error-free through the channel at a rate greater than the capacity. This is *Shannon's Second Coding Theorem* (also called the *Channel Coding Theorem* or the *Noisy Coding Theorem*). If you're sending information at a rate below the channel capacity you are wasting part of the channel and should seek a better code (Chapter 13 will look at how to do this); if you're sending information near the capacity you are doing as well as possible and there is no point in trying to improve the code; and there is no hope of reliably sending messages much above the capacity.

A few points about channel coding:

- As the transmission rate increases it might be expected that the best-case error rate will also increase; it is surprising that the error rate can remain zero until the capacity is reached (Figure 4.2).
- This proves the existence of zero-error codes but it doesn't help find them, and

of many small types of interference. Gaussian distributions are particularly important in information theory because, for a given mean and variance, they maximize the differential entropy. This makes it easy to calculate the maximum in equation (4.18). To see this, let $\mathcal{N}(x)$ be a Gaussian distribution

$$\mathcal{N}(x) = \frac{1}{\sqrt{2\pi\sigma_{\mathcal{N}}^2}} \, e^{-(x-\mu_{\mathcal{N}})^2/2\sigma_{\mathcal{N}}^2} \quad , \tag{4.19}$$

and let $p(x)$ be an arbitrary distribution with mean μ_p and variance σ_p^2. Then

$$\begin{aligned} & -\int_{-\infty}^{\infty} p(x) \ln \mathcal{N}(x)\, dx \\ &= -\int_{-\infty}^{\infty} p(x) \left[-\ln \sqrt{2\pi\sigma_{\mathcal{N}}^2} - \frac{(x-\mu_{\mathcal{N}})^2}{2\sigma_{\mathcal{N}}^2} \right] dx \\ &= \ln \sqrt{2\pi\sigma_{\mathcal{N}}^2} + \frac{\sigma_p^2 + \mu_p^2 - 2\mu_p\mu_{\mathcal{N}} + \mu_{\mathcal{N}}^2}{2\sigma_{\mathcal{N}}^2} \quad . \end{aligned} \tag{4.20}$$

This depends only on the mean and variance of $p(x)$ and so if $q(x)$ has the same mean and variance then

$$-\int_{-\infty}^{\infty} p(x) \ln \mathcal{N}(x)\, dx = -\int_{-\infty}^{\infty} q(x) \ln \mathcal{N}(x)\, dx \quad . \tag{4.21}$$

Now consider the difference in the entropy between a Gaussian distribution $\mathcal{N}$ and another one p with the same mean and variance:

$$\begin{aligned} H(\mathcal{N}) - H(p) &= -\int_{-\infty}^{\infty} \mathcal{N}(x) \ln \mathcal{N}(x)\, dx + \int_{-\infty}^{\infty} p(x) \ln p(x)\, dx \\ &= -\int_{-\infty}^{\infty} p(x) \ln \mathcal{N}(x)\, dx + \int_{-\infty}^{\infty} p(x) \ln p(x)\, dx \\ &= \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{\mathcal{N}(x)}\, dx \\ &= D(p, \mathcal{N}) \geq 0 \quad . \end{aligned} \tag{4.22}$$

The differential entropy in any other distribution will be less than that of a Gaussian with the same mean and variance. This differs from the discrete case, where the maximum entropy distribution was a constant, or an exponential if the energy is fixed.

Now return to our Gaussian channel $y = x + \eta$. Typically the input signal will be constrained to have some maximum power $S = \langle x^2 \rangle$. The capacity must be found by maximizing with respect to this constraint:

$$C = \max_{p(x):\langle x^2 \rangle \leq S} I(x, y) \quad . \tag{4.23}$$

The mutual information is

$$\begin{aligned} I(x, y) &= H(y) - H(y|x) \\ &= H(y) - H(x + \eta|x) \\ &= H(y) - H(\eta|x) \\ &= H(y) - H(\eta) \quad , \end{aligned} \tag{4.24}$$

where the last line follows because the noise is independent of the signal. The differential entropy of a Gaussian process is straightforward to calculate (Problem 4.3):

$$H(\mathcal{N}) = \frac{1}{2}\log(2\pi e N) \tag{4.25}$$

(where $N = \sigma_{\mathcal{N}}^2$ is the noise power). The mean square channel output is

$$\begin{aligned}\langle y^2\rangle &= \langle (x+\eta)^2\rangle \\ &= \langle x^2\rangle + 2\langle x\rangle \underbrace{\langle\eta\rangle}_{0} + \langle\eta^2\rangle \\ &= S + N \quad .\end{aligned} \tag{4.26}$$

Since the differential entropy of x must be bounded by that of a Gaussian process with the same variance, the mutual information will be a maximum for

$$\begin{aligned}I(x,y) &= H(y) - H(\eta) \\ &\le \frac{1}{2}\log[2\pi e(S+N)] - \frac{1}{2}\log(2\pi e N) \\ &= \frac{1}{2}\log\left(1 + \frac{S}{N}\right) \quad .\end{aligned} \tag{4.27}$$

The capacity of a Gaussian channel grows as the logarithm of the ratio of the signal power to the channel noise power.

Real channels necessarily have finite bandwidth. If a signal is sampled with a period of $1/2\Delta f$ then by the *Nyquist Theorem* the bandwidth will be Δf. If the (one-sided, white) noise power spectral density is N_0, the total energy in a time T is $N_0\Delta fT$, and the noise energy per sample is $(N_0\Delta fT)/(2\Delta fT) = N_0/2$. Similarly, if the signal power is S, the signal energy per sample is $S/2\Delta f$. This means that the capacity per sample is

$$\begin{aligned}C &= \frac{1}{2}\log\left(1 + \frac{S}{N}\right) \\ &= \frac{1}{2}\log\left(1 + \frac{S}{2\Delta f}\frac{2}{N_0}\right) \\ &= \frac{1}{2}\log_2\left(1 + \frac{S}{N_0\Delta f}\right) \ \frac{\text{bits}}{\text{sample}} \quad .\end{aligned} \tag{4.28}$$

If the signal power equals the noise power, then each samples carries 1/2 bit of information.

Since there are $2\Delta f$ samples per second the information rate is

$$\begin{aligned}C &= \Delta f \log\left(1 + \frac{S}{N}\right) \\ &= \Delta f \log_2\left(1 + \frac{S}{N_0\Delta f}\right) \ \frac{\text{bits}}{\text{second}} \quad .\end{aligned} \tag{4.29}$$

This is the most important result in this chapter: the capacity of a band-limited Gaussian channel. It increases as the bandwidth and input power increase, and decreases as the noise power increases.

This remarkable volume is astonishing in its breadth, focus, and relevance. It treats a dozen important topics, not to be found in any other single book, with wisdom and wit. Gershenfeld is our knowledgeable and thoughtful guide to the principles that underlie the far-flung world of information technology. This handy book gives you a way to connect your fragmentary knowledge, and seem much smarter than you are! It reads like a marvelous collection of excellent short stories, ultimately deeply connected in their themes. I picked it up at the end of a long day, started paging through it, and emerged, exhilarated, a couple of hours later! Ideal for professionals and dilettantes alike, this fine volume is a "must have" reference for anyone whose work (or play) involves bits and bytes, communications and computation.

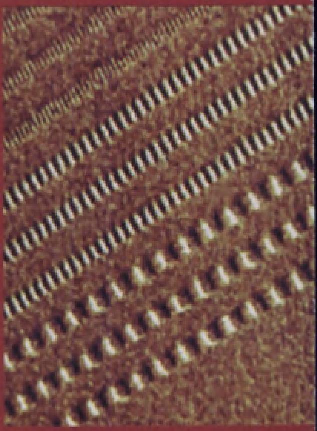
PROF. PAUL HOROWITZ
Author of *The Art of Electronics*

Information is more than abstract bits and bytes; it is by nature physical. In *The Physics of Information Technology* MIT's Neil Gershenfeld presents his eclectic, imaginative, and wide-ranging treatment of physical processes and techniques that will allow a scientist or engineer to bring together many apparently unrelated ideas in order to comprehend deeply the nature of information. This book should receive wide circulation, and I expect the resulting appreciation of the physics will serve to advance both fundamental and applied research across the range of topics he covers. It is through developing these kinds of essential connections that science realizes its great promise to enrich our economic, social, and intellectual life.

THE HONORABLE RUSH HOLT
U.S. Representative (and physicist)

Our culture is famously focused on information but only recently has much attention been given to how information flows through physical processes. This new perspective, embodied in Gershenfeld's book, has the potential to reshape physical theory at a fundamental level and stimulate profound but yet unknown mathematical abstractions. More modestly, it may lead to a technological revolution that will transform the world.

MICHAEL FREEDMAN
Fields Medalist



CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 0-521-58044-7

9 780521 580441

Cover pictures: full details of cover pictures can be found on the front endpaper.

PRINTED IN THE UNITED KINGDOM