1 Introduction

Nihil est in intellectu quod non prius fuerit in sensu.[1]

1.1 The Physics of Information

This book describes the limits for the communication of information with waves. How many ideas can we communicate by writing on a sheet of paper? How well can we hear a concert? How many details can we distinguish in an image? How much data can we get from our internet connection? These are all questions related to the transport of information by waves. Our sensing ability to capture the differences between distinct waveforms dictates the limits of the amount of information that is delivered by a propagating wave. The problem of quantifying this precisely requires a mathematical description and a physical understanding of both the propagation and the communication processes.

We focus on the propagation of electromagnetic waves as described by Maxwell's theory of electromagnetism, and on communication as described by Shannon's theory of information. Although our treatment is mostly based on classical field theory, we also consider limiting regimes where the classical theory must give way to discrete quantum formulations. The old question of whether information is physics or mathematics resounds here. Information is certainly described mathematically, but we argue that it also has a definite physical structure. The central theme of this book is that Shannon's information-theoretic limits are natural. They are revealed by observing physical quantities at certain asymptotic scales where finite dimensionality emerges and observational uncertainties are averaged out. These limits are also rigorous, and obey the mathematical rules that govern the model of reality on which the physical theories are based.

1.1.1 Shannon's Laws

Originally introduced by Claude Elwood Shannon in 1948, and continuing up to its latest developments in multi-user communication networks, information theory describes

[1] Empiricist claim adopted from the Peripatetic school. A principle subscribed to by Aristotle, St. Thomas, and Locke; opposed by Plato, St. Augustine, and Leibniz.

in the language of mathematics the limits for the communication of information. These limits are operational, of real engineering significance, and independent of the semantic aspects associated with the communication process. They arise from the following set of constraints expressing our inability to appreciate the *infinitum*:

- *Finite energy*: Communication occurs by a finite expenditure of energy.
- *Finite dimensionality*: Communication occurs by selecting among a range of possible choices, each identified by a finite number of attributes.
- *Finite resolution*: Each attribute can be observed with limited precision.

According to Shannon, reliable communication of information occurs if the probability of miscommunication *tends* to zero in appropriate limiting regimes. In these regimes, some interesting physical phenomena also occur: the space–time fields used to convey information become amenable to a discrete representation, and the finite dimensionality of the physical world is revealed. This allows us to view information-theoretic results as being imposed by the laws of nature. Information theory classically considers time asymptotics, used to describe point-to-point communication. Spatial asymptotics are their natural counterparts, used to extend the description to communication between multiple transmitters and receivers, and to the remote sensing of the world around us. There is a beautiful duality between the two, and this book attempts to capture it by providing a unified treatment of space–time waveforms.

1.1.2 Concentration Behaviors

At the basis of the asymptotic arguments leading to information-theoretic limits is the notion of *concentration*.

Consider a space–time waveform $f(x,y,z,t)$ of finite energy, transmitted for T seconds. As $T \to \infty$, we can define the effective frequency bandwidth of the waveform as the effective spectral support in the Fourier-transformed angular frequency domain – see Figure 1.1. This definition is made possible by the mathematics at the basis of wave theory that predict *spectral concentration*. As the time domain support is stretched, the signal, when viewed in the frequency domain, can be more and more concentrated inside the bandwidth. Thanks to this phenomenon, electromagnetic signals can be considered, for large T, as occupying an essentially finite bandwidth. Signals of finite energy and finite bandwidth enjoy another important mathematical property. They exhibit a limit on the amount of variation they can undergo in any given time interval and thus, when viewed at finite resolution, on the amount of information they can carry over time. The same limitation also applies to the spatial domain. As the region where the signal is observed is stretched by scaling all of its coordinates, spectral concentration occurs, and this allows the definition of the effective bandwidth in the wavenumber domain, that is the Fourier transform of the spatial domain. This limits the number of spatial configurations of the waveform, and thus, when viewed at finite resolution,

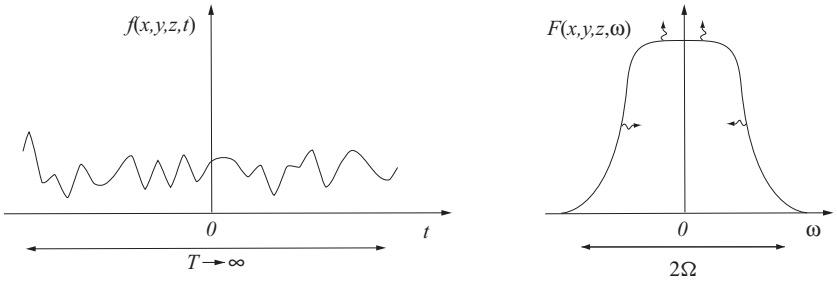

Fig. 1.1 Spectral concentration.

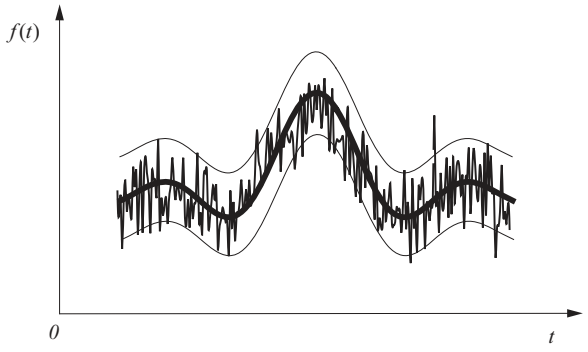

Fig. 1.2 Probabilistic concentration.

the amount of information it can carry over space. This limitation is important in the context of network information theory, when multiple transmitters and receivers in a communication system are distributed in space. It is also important in the context of imaging systems, where it leads to spatial resolution limits of the constructed image.

When considering space and time asymptotics, another kind of concentration phenomenon also occurs. The precision level at which the signal can be observed *probabilistically concentrates* around its typical value. Every physical apparatus measuring a signal is affected by a measurement error: repeated measurements appear to fluctuate randomly by a small amount. This is a consequence of the quantized nature of the world observed at the microscopic scale. Over many repetitions, the uncertainty with which the signal is observed is typically contained within its standard deviation – see Figure 1.2. This allows us to view the uncertainty of the observation as concentrated around its typical value and determines a resolution limit at which the signal can be observed. Combined with the constraints on the form of the signal due to spectral concentration mentioned above, it poses an ultimate limit on the amount of information that can be transported by waves in time and space.

The same concentration behaviors leading to information-theoretic limits are also at the basis of quantum mechanics and statistical mechanics. Spectral concentration is at the basis of Heisenberg's uncertainty principle, stating that physical quantities related by

Fourier transforms cannot be determined simultaneously, as pinpointing one precisely always implies the smearing of the other. On the other hand, probabilistic concentration is at the heart of statistical mechanics. Due to probabilistic concentration, only some realizations of a stochastic process have non-negligible probability of occurrence, and these typical outcomes are nearly equally probable. Based on this premise, statistical mechanics explains the thermodynamic behavior of large systems in terms of typical outcomes of random microscopic events. The information-theoretic approach is another instance of this method, as it exploits probabilistic concentration to describe the typical behavior of encoding and decoding systems formed by large ensembles of random variables.

1.1.3 Applications

When discussing information carried by electromagnetic waves, it is also natural to mention practical applications. In the context of electromagnetics, the operational aspects of information theory have found vast applications ranging from remote sensing and imaging to communication systems. During the wireless revolution of the turn of the millennium, the physical layer of digital communications has successfully been abstracted using approximate probabilistic models of the propagation medium. Information theory has fruitfully developed in this framework, and important theoretical contributions have inspired creative minds, who improved engineering designs. Many practical advancements have been influenced by a clearer understanding of the information-theoretic limits of different channel models. As theoretical studies found practical applications, industry flourished. Sometimes, however, this has come at the expense of casting a shadow on the physical limits of communication. The fundamental question posed by Shannon regarding the ultimate limits of communication has been obfuscated by the myriad results regarding different approximate models of physical reality.

Research in wireless communication has expanded the knowledge tree into an intricate forest of narrow cases, more of interest to the practitioner than to the scientist seeking a fundamental understanding. These cases have provided useful design guidelines, but they have also somewhat hidden the fundamental limits. This situation is somehow natural: as a field becomes more mature, improvements tend to become more sectorized, and technology, rather than basic advancements, becomes the main driver for progress. Maturity, however, should also open up the opportunity of revisiting, reordering, reinterpreting, and pruning the knowledge tree, revealing its basic skeleton. With it, we also wish to reveal the misconception that a rigorous physical treatment is too complex to provide valuable engineering insights. A physical treatment not only shows that Shannon's theory is part of the fundamental description of our physical universe, but it also provides insights into the design and operation of real engineering systems. After all, if, as Alfréd Rényi (1984) put it, information theory came out of the realization that the flow of information, like other physical quantities, can be expressed numerically, then our engineering designs should best exploit its physical nature.

For this reason, we devote Chapter 7 to discussing communication technologies, and we revisit these technologies in the light of our physical treatment of information. Of course, we can only discuss some of the fundamental principles; for a more in-depth perspective the reader should refer to the wide range of available engineering literature.

1.2 The Dimensionality of the Space

To quantify the amount of information carried by electromagnetic waves of finite energy, we represent the possible messages that the transmitter may send by an ensemble of waveforms, and then attempt to quantify the amount of information transferred by the correct selection at the receiver of one element from this ensemble, according to Shannon's theory. This selection implies that a certain amount of information has been transported between the two. To realize this program, we are faced with a first fundamental question:

> How many distinct waveforms can we possibly select from?

The answer depends on the size of the signals' space, or the *number of degrees of freedom*, available for communication and on the resolution at which each degree of freedom can be observed. The number of degrees of freedom is limited by the nature of the wave propagation process, while the resolution is limited by the uncertainty associated with the observation process.

1.2.1 Bandlimitation Filtering

Any radiated waveform has an essentially finite bandwidth. Due to the interaction with the propagation medium and with the measurement apparatus used to detect the signal, the high-frequency components are cut off, and any signal appears as the output of a linear filter. Typically this filter also shapes the frequency profile, distorting the transmitted waveform. A simple example of this phenomenon occurs in a scattering environment. Due to multiple scattering, multiple copies of the transmitted waveform, carrying different delays and attenuations while traveling on different paths, may overlap at the receiver, creating interference and distorting the original signal – see Figure 1.3. While in ideal free space the magnitude of the frequency response of the propagation environment would be flat across all frequencies, and the phase of the response would be proportional to the distance between transmitter and receiver, in the presence of a large amount of scattering the response is a highly varying signal, due to the interference over the multiple scattered paths.

An analogous effect occurs in the spatial domain. The signal observed along any spatial cut-set that separates transmitters and receivers is the image of all the elements radiating from one side of the cut to the other. These radiating waveforms may interfere along the cut, leading to spatial filtering – see Figure 1.4.

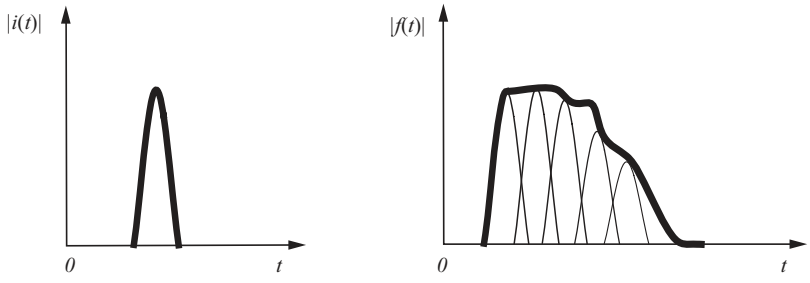

Fig. 1.3 Spreading of the signal in the time domain due to multiple scattering. Transmitted signal: $i(t)$; received signal: $f(t)$.

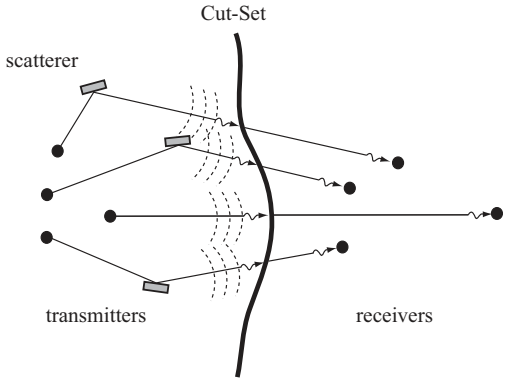

Fig. 1.4 Multiple signals overlap over the cut-set boundary.

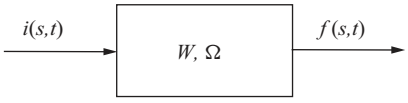

Fig. 1.5 Propagation filtering.

A block diagram of the propagation filtering effect that occurs when a scalar source current $i(s,t)$ of one spatial and one temporal variable produces a scalar electromagnetic field $f(s,t)$ observed along a given cut-set boundary is depicted in Figure 1.5. The figure shows that the effect of propagation is analogous to that of a linear filter of frequency cut-off Ω and of wavenumber cut-off W. The form of the transfer function depends on the features of the environment where propagation occurs and is studied in Chapters 8 and 9. This filtering operation limits the number of distinct space–time waveforms that can be observed at the receiver, and makes the space of electromagnetic signals essentially bandlimited and suitable for an information-theoretic analysis.

1.2.2 The Number of Degrees of Freedom

The physics of propagation dictate that any observed electromagnetic field is an essentially bandlimited function. This basic property allows us to define the size of the signals' space in terms of the number of degrees of freedom. Consider a one-dimensional, real, scalar waveform f of a single scalar variable t. We assume that f is square-integrable, and

$$\int_{-\infty}^{\infty} f^2(t)dt \leq E. \tag{1.1}$$

This ensures that the waveform can be expanded in a series of, possibly complex, orthonormal basis functions $\{\psi_n\}$,

$$f(t) = \sum_{n=1}^{\infty} a_n \psi_n(t), \tag{1.2}$$

where

$$a_n = \int_{-\infty}^{\infty} f(t)\psi_n^*(t)dt. \tag{1.3}$$

The equality in (1.2) is intended in the "energy" sense:

$$\lim_{N\to\infty} \int_{-\infty}^{\infty} [f(t) - f_N(t)]^2 dt = 0, \tag{1.4}$$

where

$$f_N(t) = \sum_{n=1}^{N} a_n \psi_n(t). \tag{1.5}$$

In the language of mathematics, f is in $L^2(-\infty,\infty)$, and it can be viewed as a point in an infinite-dimensional space of coordinates given by the coefficients $\{a_n\}$ in (1.3). By varying the values of these coefficients, we can create distinct waveforms and use them to communicate information. If the orthonormal set of basis functions $\{\psi_n\}$ is complete, then using (1.2) we can construct any element in the space of signals defined by (1.1). By associating a waveform in this space with a given message that the transmitter wishes to communicate, the correct selection of the same waveform at the receiver implies that a certain amount of information is transferred between the two. One may reasonably expect that only a finite number of coefficients is in practice needed to specify the waveform up to any given accuracy, while using a larger number does not significantly improve the resolution at the receiver. It turns out that the question of what the smallest N is beyond which varying higher-order coefficients does not change the form of the waveform significantly has a remarkably precise answer.

Determining this number over all possible choices of basis functions is a question first posed by Kolmogorov in 1936, and corresponds to determining the *number of degrees of freedom* of the waveform. Consider an observation interval $[-T/2, T/2]$, and introduce the norm

$$\|f\| = \left(\int_{-T/2}^{T/2} f^2(t)dt\right)^{1/2}. \tag{1.6}$$

The problem amounts to determining the interval of values of N for which the approximation error for any signal f,

$$\|e_N\| = \|f(t) - f_N(t)\|, \tag{1.7}$$

transitions from being close to its maximum to being close to zero.

For signals of spectral support in an interval of size 2Ω, and observed over a time interval of size T, the angular frequency bandwidth Ω and the size of the observation interval play a key role in determining the number of degrees of freedom, as the approximation error undergoes a *phase transition* at the scale of the product ΩT.

For any $\epsilon > 0$, letting N_ϵ be the minimum number of basis functions for which the normalized energy of the error

$$\frac{\|e_{N_\epsilon}\|^2}{E} \leq \epsilon, \tag{1.8}$$

and letting

$$N_0 = \frac{\Omega T}{\pi}, \tag{1.9}$$

we have

$$\lim_{N_0 \to \infty} \frac{N_\epsilon}{N_0} = 1. \tag{1.10}$$

This result was a crowning achievement of three scientists, Henry Landau, Henry Pollak, and David Slepian, working at Bell Laboratories in the 1960s and 70s, and we discuss it in detail in Chapters 2 and 3. It identifies the number of degrees of freedom with the time-bandwidth product $N_0 = \Omega T/\pi$, and shows that this number is, up to first order, independent of the level of approximation ϵ. This is evident by rewriting (1.10) as

$$N_\epsilon = N_0 + o(N_0) \quad \text{as } N_0 \to \infty, \tag{1.11}$$

where the dependence on ϵ appears hidden as a pre-constant of the second-order term $o(N_0)$ in the phase transition of the number of degrees of freedom. Varying the approximation level does not affect the first-order scaling of the result. Figure 1.6 shows the transition of the approximation error for the optimal choice of basis functions. According to (1.10), the transition occurs in a small interval and tends to become a step function when viewed at the scale of N_0.

A widely used representation for bandlimited waveforms is the cardinal series that uses sampled values of the waveform as coefficients $\{a_n\}$ and real functions of the form $\text{sinc}\,(t) = (\sin t)/t$ as the basis set $\{\psi_n\}$, yielding the Kotelnikov–Shannon–Whittaker sampling representation

$$f(t) = \sum_{n=-\infty}^{\infty} f(n\pi/\Omega)\,\text{sinc}\,(\Omega t - n\pi), \tag{1.12}$$

which interpolates the signal from regularly spaced samples at frequency Ω/π. This representation is suboptimal in terms of the approximation error (1.7). The optimal interpolating functions, called prolate spheroidal wave functions, that achieve the

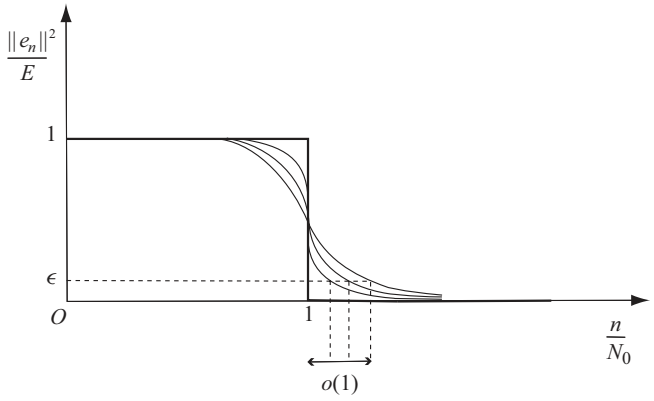

Fig. 1.6 Phase transition of the approximation error. The transition becomes sharper, and its width, viewed at the scale of N_0, shrinks to zero as $N_0 \to \infty$.

smallest approximation error are obtained by solving an eigenvalue problem that we study in Chapters 2 and 3.

1.2.3 Space–Time Fields

The electromagnetic field is in general a function of four scalar variables: three spatial and one temporal. It follows that in order to appreciate the total field's informational content in terms of degrees of freedom, we need to extend the treatment above to higher dimensions.

Let us first consider the canonical case of a two-dimensional domain of cylindrical symmetry, in which an electromagnetic field is radiated by current sources located inside a circular domain of radius r, and oriented perpendicular to the domain. The sources can also be induced by multiple scattering inside the domain. In any case, the radiated field away from the sources is completely determined by the field on the cut-set boundary surrounding the sources and through which it propagates – see Figure 1.7. On this boundary, we can refer to a scalar field $f(\phi,t)$ that is a function of only two scalar variables: one angular and one temporal. The corresponding four representations, linked by Fourier transforms, are depicted in Figure 1.8, where ω indicates the transformed coordinate of the time variable t and w indicates the wavenumber that is the transformed coordinate of the angular variable ϕ.

Letting Ω be the angular frequency bandwidth and W be the wavenumber bandwidth, we now wish to determine the total number of degrees of freedom of the space–time field $f(\phi,t)$. To visualize the phase transition, we fix the bandwidth Ω and the size of the angular observation interval $S = 2\pi$, and scale the time support where the signal is observed $T \to \infty$ and the wavenumber bandwidth $W \to \infty$. Using the results of the monodimensional case, we have that as $T \to \infty$ the number of time–frequency degrees of freedom is of the order of

$$N_0 = \frac{\Omega T}{\pi}. \tag{1.13}$$

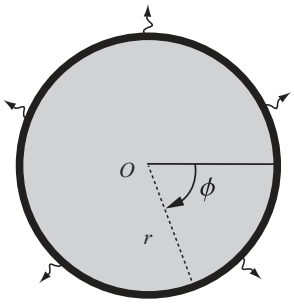

Fig. 1.7 Cylindrical propagation, cut-set boundary.

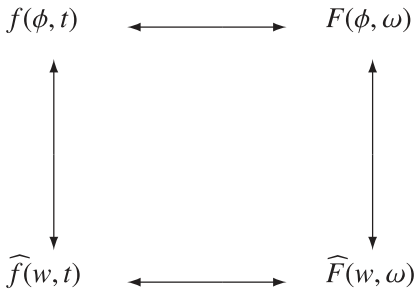

Fig. 1.8 Four field representations linked by Fourier transforms.

In a symmetric fashion, letting the wavenumber bandwidth $W \to \infty$, we have that the number of space–wavenumber degrees of freedom over an observation interval $S = 2\pi$ is of the order of

$$N_0 = \frac{W 2\pi}{\pi}. \tag{1.14}$$

The wavenumber bandwidth is related to the frequency of transmission. As we shall see in Chapter 8, every positive frequency component $\omega > 0$ of the signal has a wavenumber bandwidth that is, for any possible configuration of sources and scatterers inside the circular radiating domain, at most $W = \omega r/c$. It follows that the appropriate asymptotic regime $W \to \infty$ can be obtained by letting $r \to \infty$, so that by (1.14) the number of space–wavenumber degrees of freedom at angular frequency ω becomes of the order of

$$N_0(\omega) = \frac{2\pi r\omega}{c\pi} = \frac{2\pi r}{\lambda/2}, \tag{1.15}$$

where $\lambda = 2\pi c/\omega$ is the radiated wavelength, and c is the propagation speed of the field. The total number of degrees of freedom can now be obtained by integrating (1.15) over the bandwidth, and multiplying the result by T/π. It follows that as $r, T \to \infty$ the total number of degrees of freedom is of the order of

$$N_0 = \frac{2\pi rT}{c\pi^2} \int_0^{\Omega} \omega d\omega = \frac{2\pi rT\Omega^2}{2c\pi^2}. \tag{1.16}$$