Shannon's Sampling Theorem

Maurice Dodson

The Sampling Theorem is one of the key results in communication theory, giving a representation of a bandlimited analogue signal as a sum of terms involving the values (samples) of the signal taken at the Nyquist rate (twice the maximal frequency of the signal). In his fundamental and definitive framework for information theory, Shannon used the Sampling Theorem to establish the theoretical equivalence of analogue and digital signals. Later with the advent of extremely fast digital computers, the theorem served as a basis of efficient practical techniques for digital/analogue conversion, vital in modern communication systems. In essence, the theorem was first proved in interpolation theory and is closely related to a wide range of other results in mathematics. The sampling rate specified in the Sampling Theorem is crucial and for lowpass signals must exceed the Nyquist rate in order to prevent 'aliasing'; higher rates also lead to smaller errors in digital/analogue conversion. For multiband signals with spectra consisting of more than one frequency band, lower sampling rates which still prevent aliasing can be achieved if chosen appropriately.

Introduction

MODERN communication systems depend upon the equivalence of continuous and digital signals. This might appear surprising since it would seem that a continuous signal which has a value for each real number could always contain more information than a discrete signal with values limited to discrete points. The key to this equivalence is the Sampling Theorem, widely associated with Claude Shannon[1,2] who laid down the foundations of information theory in three classic papers published in 1948–49. A recent article on Shannon paints an attractive picture of an inventive and irreverent yet modest character[3].

The significance of Shannon's ideas was recognized by both engineers and mathematicians who placed his work on a more secure and rigorous footing (see, for example ref. 4). To set the context of the Sampling Theorem, we will begin with a brief description of a general representation of a communication system. Further details can be found in any introductory account of information theory. Shannon's original 1948 papers, which are clear and accessible, are as good a source as any and are in the book[5] by Shannon and Weaver but a more up-to-date account is given in ref. 6.

Communication systems and signals

Shannon's analysis began with a very general definition of a communication system, which he described as consisting of five elements:

(1) An information source. This might be a voice, a physical reading such as temperature or voltage, or a keyboard.
(2) A transmitter. This converts the message to a signal suitable for transmission. A voice might be converted into electrical impulses and transmitted by wire or radio.
(3) A channel. This is the medium of the message and might be a wire or a band of frequencies.
(4) The receiver. This receives the signal and converts it back to the original message, in practice never exactly.
(5) The destination. The person (e.g. listener) or device for which the message was intended.

The information from the source might consist of a sequence of discrete symbols, such as letters from an alphabet, or might be a continuous reading such as a voltage, temperature or pressure. A continuous signal is called *analogue*. The means of transmitting the message can also be discrete (e.g. semaphore) or analogue (e.g. a current in a wire), though in practice it is usually the latter.

In his analysis, Shannon used a quantitative measure of information derived from the observation that a discrete message can be regarded as a selection from a finite set of symbols. For instance an English word is made up from an alphabet of 26 letters (plus a space). More simply but just as generally a message can be considered to be made up of a string of 0's and 1's, or *bits* (from binary digits). These were used to make precise the idea of information content and to establish

Maurice Dodson is in the Department of Mathematics, University of York, Heslington, York YO1 5DD, UK.

fundamental results on redundancy, the 'entropy' of a message, source structure, noise and channel capacity. It is worth mentioning Shannon's counterintuitive second fundamental theorem that noise limits the rate of transmission but not the accuracy of a message. The references cited above give more details.

Discrete and analogue signals

Although the theory Shannon developed applied to discrete sequences essentially consisting of a stream of 0's and 1's, most signals are analogue. For example, since they are continuous, speech, radio and television transmissions are analogue signals. Analogue signals are made up of waves of various frequencies, for example sound consists of compression waves and radio and television of electromagnetic waves. The frequencies and their amplitudes which make up a signal are called the *spectrum*. In the case of adult male speech the frequencies in the spectrum are in the range of 0 to 8000 cycles per second (or 8 kH). Communication engineers take signals to be real square-integrable functions (i.e. in mathematical terminology, they lie in $L^2(\mathbb{R})$, where $\mathbb{R}$ denotes the set of real numbers), corresponding to their having finite energy[7]. This places the study of signals firmly into the realm of Fourier analysis, for instance the Fourier transform of a signal is its spectrum. We shall be concerned with the classical theory of analogue or continuous deterministic signals, which are a subspace of $L^2(\mathbb{R})$. Non-deterministic signals such as noise can still be treated within a Fourier analysis setting but we shall not say much about this important development (pioneered by Norbert Wiener).

Although natural and highly effective, modelling analogue signals using $L^2(\mathbb{R})$ and Fourier analysis carries a paradox within it. The frequency components (assumed to be sinusoidal) of a physical signal are generated by vocal chords (for speech), a string (for a violin) or an oscillator (for an electrical signal) which cannot vibrate above some finite limit. Thus it seems reasonable to assume that the frequencies of analogue signals are bounded. Engineers call such signals *bandlimited* and assume that they have a maximum frequency (strictly speaking the bounded spectrum will only have a supremum or least upper bound rather than a maximum but this distinction is irrelevant in practice). On the other hand, it is equally reasonable to assume that signals start and stop within a finite time, i.e. are *timelimited* and usually engineers treat signals as being limited to times t with $|t| < T/2$ and with no frequencies greater than W. However it turns out that in the Fourier analysis framework, the only signal which is both bandlimited and timelimited is the zero signal. Although the Fourier analysis framework is used in an essential way in sampling and other areas, this paradox does not seem to cause any trouble. Later we will touch on ways of resolving it.

In order to extend his results for (time) discrete signals to analogue ones, Shannon assumed that bandlimited analogue signals can be regarded as approximately timelimited as well and used the Sampling Theorem to deduce that communicating information can in effect be treated as a discrete process. This is one of the key results in Shannon's paper[2] since it enabled him to analyse analogue signals in terms of approximately equivalent discrete signals which formed a finite dimensional space.

The Sampling Theorem

The Sampling Theorem tells us that an analogue signal with maximum frequency W is determined by its values (samples) taken every $1/2W$ sec. In Shannon's words this was 'a fact which is common knowledge in the communication art', a revealing comment on the mathematical state of the subject at that time. The intuitive justification was that an analogue signal with maximum frequency W cannot vary substantially in a time interval less than one half a cycle, i.e. in $1/2W$ sec. This had been recognized since the 1920s but the Sampling Theorem establishes more. It provides a *formula which expresses a signal in terms of the discrete* values or samples taken every $1/2W$ sec, where W is the maximum frequency of the signal. The formula for a signal $f(t)$ is:

$$f(t) = \sum_{k \in Z} f\left(\frac{k}{2W}\right) \frac{\sin \pi(2Wt - k)}{\pi(2Wt - k)}, \qquad (1)$$

where $\mathbb{Z}$ is the set of integers. Evidently the signal $f(t)$ at the time t can be reconstructed from the values

$$\cdots, f\left(\frac{-1}{W}\right), f\left(\frac{-1}{2W}\right), f(0), f\left(\frac{1}{2W}\right), f\left(\frac{1}{W}\right), f\left(\frac{3}{2W}\right), \cdots$$

of the signal at discrete points $\cdots, -1/2W, -1/W, 0, 1/2W, 1/W, 3/2W, \cdots$. This corresponds to a sampling rate of $2W$, twice the maximum frequency of the signal, and is often called the *Nyquist rate*. The formula (1), which was already known in mathematics (see next section), can be regarded as a sum of translated and weighted functions of the type

$$\frac{\sin 2\pi Wt}{2\pi Wt},$$

which is 1 when $t = 0$ and vanishes when $t = k/2W$, where k is a non-zero integer (see Figure 1).

In fact the practical significance of (1) for radio communication was discovered in 1933 by a Russian,

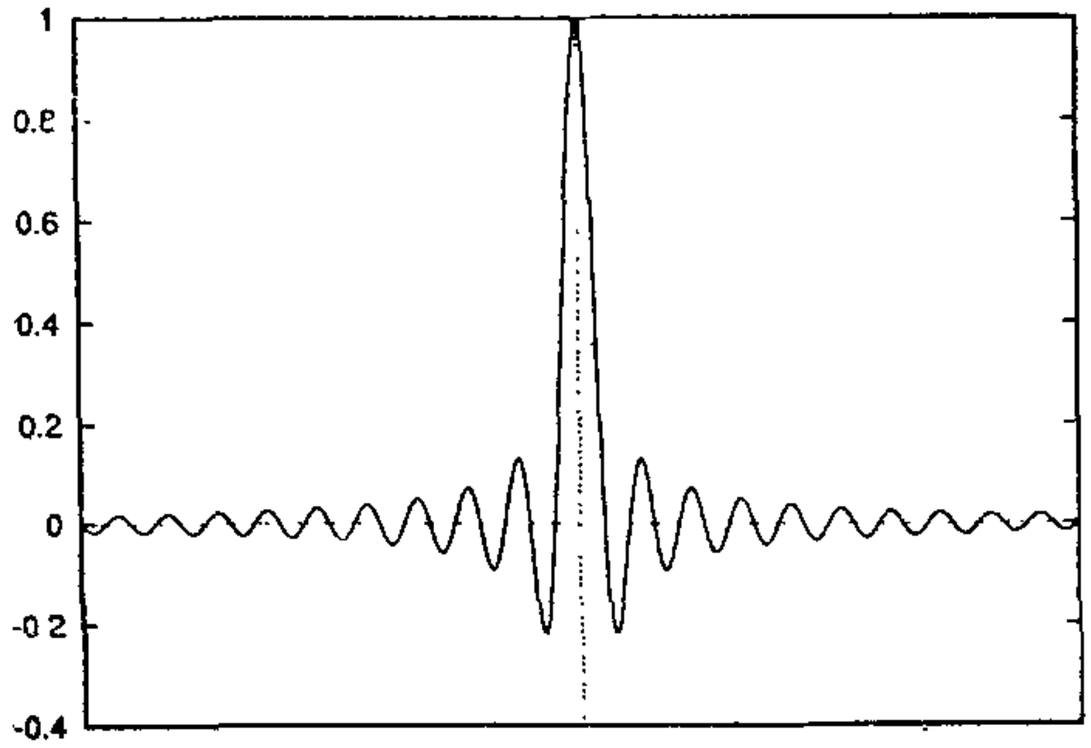

Figure 1. The graph of $\sin(2\pi Wt/2\pi Wt)$; the zeros are at the points $k/2W$, $k \neq 0$.

V. A. Kotel'nikov[8] (who retired in 1990). In Eastern Europe and the former USSR, the Sampling Theorem is named after him. A little later, in 1939, a German, Raabe[9] published a paper on the Sampling Theorem in communication. It is remarkable that two other papers on its application to communication theory were published independently in the same year (1949) as Shannon's paper (which was submitted for publication in 1940 but did not appear until after World War II)—one by a Briton, Weston[10], and the other by a Japanese, I. Someya[11]. Thus although he was by no means the first to prove the theorem (he himself gave a reference to a book[12] on interpolation theory written by J. M. Whittaker, a son of E. T. Whittaker), Shannon was the first to appreciate the general significance of the Sampling Theorem for information theory. Consequently in the West the result is usually known by his name, though the name Whittaker–Kotel'nikov–Shannon Theorem is becoming more common.

It should be pointed out that there is a drawback to the formula (1). The function f given by (1) can be extended to the complex plane by replacing t with the complex number $t+iu$. It turns out that this extended complex function is entire, i.e. analytic throughout the complex plane. Analyticity is a 'rigid' property in the sense that an analytic function is determined by its values on a set with an accumulation point, such as an interval. In particular a timelimited and bandlimited signal must be identically zero since such a signal vanishes for all sufficiently large times and hence for all time. This is why the only simultaneously time and bandlimited signal is the zero signal. This paradox is closely related to Heisenberg's uncertainty principle in quantum mechanics[13–15].

One way out of this dilemma is to drop the assumption that the signal is square-integrable and bandlimited and assume instead that the signal and its spectrum (Fourier transform) are integrable. It then follows that an error term must be added to equation (1) to give an asymptotic formula[16]. Another approach is to accept that a bandlimited signal cannot vanish outside any finite interval (although it will be negligible for times outside the interval $(-T/2, T/2)$ when T is large) and to work with signals which are concentrated on $(-T/2, T/2)$. Indeed since physical measurements of a signal are never exact and its mathematical properties, such as continuity, cannot be settled by observation, this is very reasonable. It is a remarkable fact that the *prolate spheroidal wave functions* (suitably scaled) form a complete orthonormal basis for $L^2(\mathbb{R})$ and are concentrated on $(-T/2, T/2)$. Moreover their restrictions to $(-T/2, T/2)$ are still orthogonal and complete for the space of signals time limited to $(-T/2, T/2)$. A full discussion, which involves an appreciation of the relationship between physical reality and mathematical models, is given in refs. 7, 13, 14 and will not be pursued here.

As well as the Sampling Theorem, the paper[2] contains a wealth of imaginative geometric ideas applied to signals of finite time duration T. As has been just pointed out, a signal cannot in fact be simultaneously of finite duration T and have spectrum bounded by W say. Nevertheless, Shannon observed that this approximation allows a signal to be specified by $2TW$ values and went on to develop an illuminating geometrical approach to communication theory by representing a given signal as a point in a $2TW$-dimensional vector space of all possible signals. From this geometric point of view, the transmitter is then a mapping from the space of messages to the approximately $2TW$ dimensional signal space and the receiver a mapping from signal space back to the message space. Shannon pictured the effects of noise as producing distorted messages lying inside a sphere centred at the point representing the received message. There is a rigorous version of the '$2TW$' theorem which uses the prolate spheroidal wave functions to make Shannon's approximate statement precise; the space is of dimension $2TW + o(TW)$ for large T (refs. 13, 14).

Later, Shannon[2] even raised the possibility of the efficient storing of messages by means of space-filling curves, usually regarded as a mathematical oddity of little practical significance. However, as Shannon himself pointed out, very slight errors in storing the message could result in major errors in its reconstruction. He also observed that the samples do not have to be taken at regular intervals and that a signal can be determined by choices other than its own values. For instance, the values of the signal and its derivative taken at half the Nyquist rate would serve.

With the advent of extremely fast digital computing in the seventies, the Sampling Theorem has assumed an even greater importance in applications. Besides being fundamental in information theory, the Sampling

Theorem has applications in many other areas, including communications engineering, optics, spectroscopy and prediction theory (an extensive list is given in the comprehensive survey article[17]), and above all in electronics and signal processing, where it underlies pulse amplitude and code modulation techniques for analogue/digital signal conversion. There is a nice application to X-ray crystallography, where the properties of reciprocal space are reflected in the quantity $1/2W$, being the reciprocal of the length of the interval $(-W, W)$ (ref. 18).

In applied mathematics, the Sampling Theorem is used in a variety of problems, including nuclear scattering, heat transfer and general discrete transforms (ref. 17, §D). Applications in pure mathematics, such as in interpolation theory, are to be expected and will be discussed in the next section. However this is an appropriate point to mention that there is a version of the Sampling Theorem mentioned by Shannon[2] in which the samples are the values $f(k/W)$ and the values $f'(k/W)$ of the derivative at the points k/W, $k \in \mathbb{Z}$. This result has applications in analysis and number theory[19].

There are stochastic versions of the Sampling Theorem which apply to non-deterministic signals[20,21]; Lloyd[21] gives a more general result in which the interval $(-W, W)$ is replaced by a set A satisfying a disjoint translates condition.

Mathematical origins of the Sampling Theorem

The Sampling Theorem has a tangled history going back to the late nineteenth century, a decade after Hertz had produced radio waves in the laboratory (1887). Neither Hertz's successful experiments nor Maxwell's earlier prediction in 1877 of radio waves on the basis of his equations for electromagnetism appear to have had any influence on the original discovery, in mathematics, of the Sampling Theorem. It appears to have been discovered first by a French mathematician, Borel, in 1897 and independently later in 1910 by a Briton, Whipple, who never published it. Other famous names associated with the result are Hadamard and de la Vallée Poussin; more details are given in a very readable survey article[22]. The first published proof was by Whittaker[23] in 1915, from the point of view of the mathematical theory of interpolation. In the Sampling Theorem, the object is to recover or reconstruct a signal or function with known maximum frequency from samples or values. On the other hand, in interpolation there is generally not such a restriction in the particular type (such as polynomials, analytic functions, etc.) of a function which takes on given values or data at specified points. Whittaker produced the function (1) which interpolated given values at equally spaced points and which was free of 'rapid oscillations'. This function, called the *cardinal series*, is as has been pointed out analytic and so very well behaved. It is interesting that over 75 years ago Whittaker thought that because of its nice properties, the cardinal series might be of use to applied mathematicians who

> for long past have complained that pure mathematics is daily becoming more complicated and harder to understand. This complaint refers chiefly to the increased rigour with which the theories of analysis are now expounded, and which is closely connected with the extension of knowledge regarding discontinuities, singularities, and other phenomena of which the older mathematics took no account. Indeed, the modern theory of functions of a real variable is concerned largely with cases in which the distribution of fluctuations and singularities transcends all intuitive or geometrical representation.

The theorem has connections with other areas of mathematical analysis, above all in Fourier analysis and complex function theory. For example, one half of the Paley–Wiener Theorem asserts that if the Fourier transform $\hat{f}$ of a function $f: \mathbb{R} \to \mathbb{C}$ vanishes for $|w| > W$ then f can be extended to an entire function (i.e. one which has no poles in the complex plane $\mathbb{C}$). Hardy[24] called these *Paley–Wiener* functions and showed that they formed a subspace (a *Paley–Wiener space*) of the Hilbert space $L^2(\mathbb{R})$ of square integrable functions. This space plays an important part in making some of Shannon's ideas rigorous (see refs. 13, 14). By their very nature, Paley–Wiener functions can be regarded as bandlimited analogue signals and so have a representation by Whittaker's cardinal series as in (1).

The results hinge on the family

$$\frac{\sin \pi(2Wt-k)}{\pi(2Wt-k)}, k \in \mathbb{Z},$$

being a complete orthonormal basis for the subspace. This relates the Sampling Theorem to the Riesz–Fischer Theorem since the Sampling Theorem can be interpreted as saying that the space $L^2([-1/2, 1/2])$ of square integrable functions on $[-1/2, 1/2]$ is isometric to the space ℓ^2 of square summable sequences. The Sampling Theorem is also equivalent to some fundamental results in analysis, such as the Poisson summation formula, the Cauchy integral formula and the duality between the circle group and the group of integers. Consequently the Sampling Theorem is related to many classical results in mathematics[16,25]. However it should be pointed out that the slow rate of convergence of the Cardinal series in (1) makes the Sampling Theorem less useful in numerical analysis.

The proof of the Sampling Theorem

The proof in Shannon's paper[2] relies on the fact that a periodic function g with period $2W$ has a representation as a Fourier series, i.e. as an infinite sum of the form

$$g(x)=\sum_{k\in Z} c_k e^{\pi i k x/W}. \tag{2}$$

The constants c_k depend on the function g and are called the *Fourier coefficients* of g. The essence of the proof is that the signal shown in Figure 2*a* has spectrum confined to the interval $(-W, W)$ (Figure 2*b*). This spectrum is repeated on the intervals $(2kW-W, 2kW+W)$ for $k=\pm1, \pm2, \cdots$ along the real axis (see Figure 2*c*).

Another way of looking at this is to think of the spectrum being translated by $2W$ again and again to the left and right, filling the whole line with (non-overlapping) copies of the spectrum. This spectrum repetition in the frequency domain gives a periodic function which can be represented by a Fourier series and so determined by the Fourier coefficients $c_k=f(k/2W)$ in the series (2) (see Figure 2*d*).

Now since the periodic function can be reconstructed from a knowledge of the coefficients c_k, these coefficients can equally be regarded as determining the periodic function. In other words in Figure 2, you can go from (*d*) to (*c*). It turns out that the discrete Fourier coefficients for g in (2) are given by $c_k=f(k/2W)$ (the denominator in $f(k/2W)$ arises from the period of g being $2W$ instead of the more usual 1). Hence the c_k lie on the original signal which can thus be recovered by removing the repetitions by filtering out the added frequencies to regain the original spectrum. In other words, in Figure 2 you can go from (*d*) back to (*a*) via (*c*) and (*b*), and so recover the original signal from the samples.

Another way of looking at this is to regard the sequence of samples in (*d*) as in effect a stream of impulses weighted by the sample values $f(k/2W)$ which is passed through the filter which removes frequencies above W to give the value of $f(t)$ of the signal at the time t. This process can be expressed by the formula

$$f(t)=\sum_{k=-\infty}^{\infty} f(k/2W)\delta(t-k/2W)\star\frac{\sin 2\pi Wt}{2\pi Wt}. \tag{3}$$

(Interpreted appropriately, this is essentially the same equation as (1).)

Implementation

Although engineering in spirit, the proof that Shannon gave cannot be implemented exactly in practice for analogue signals. There are two reasons for this. First, the samples (Fourier coefficients) in the theory are perfect impulses (weighted delta functions) which do not exist physically. Secondly, 'quantization' errors associated with digitizing the sample values cannot be avoided, although they can be made very small. They arise when a signal value such as a voltage can only be measured to, say, the nearest 1/10th of a volt. Nevertheless despite these limitations, there is an effective and widely used technique based on Shannon's proof for reconstructing analogue signals.

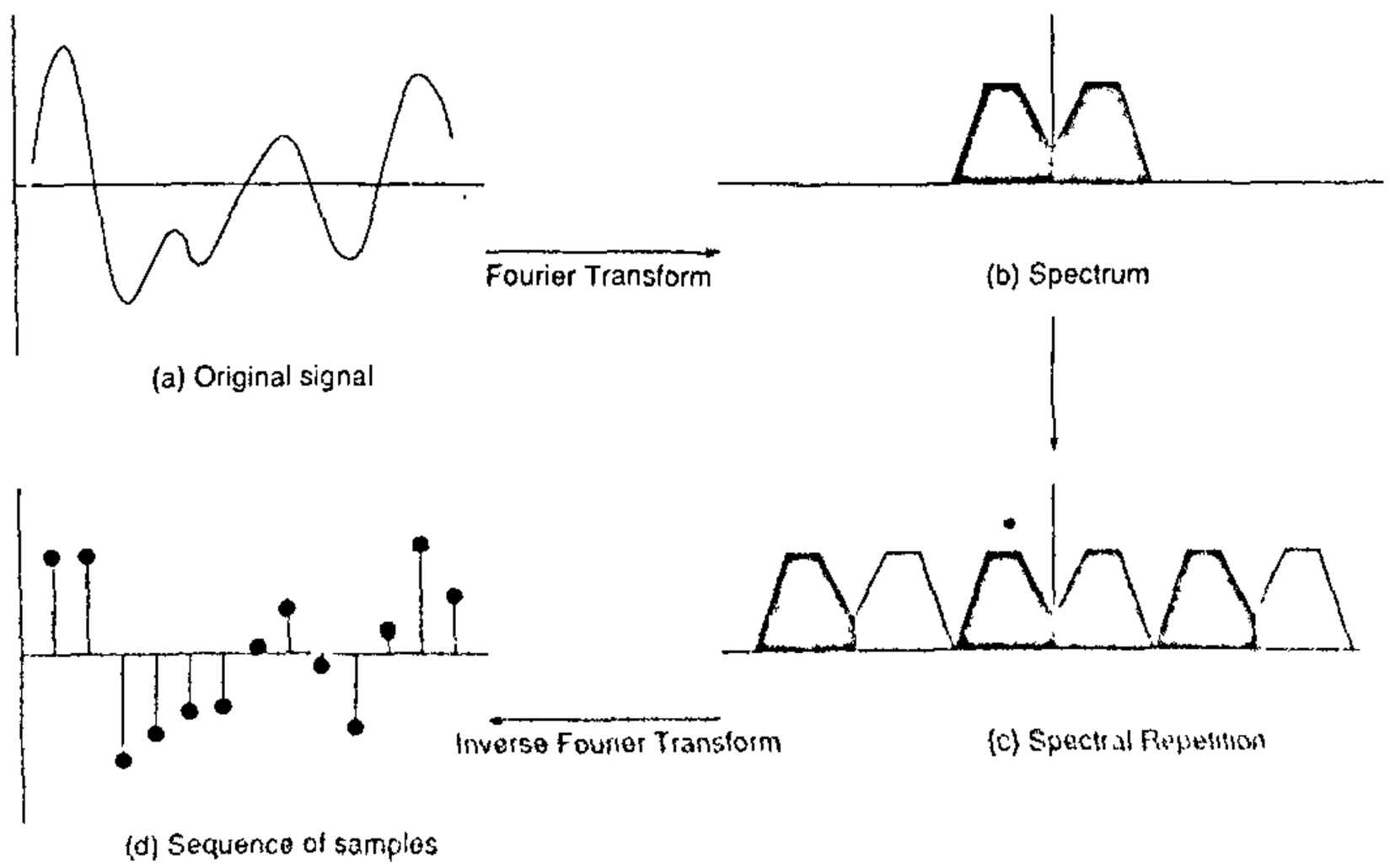

Figure 2. A diagrammatic proof of the Shannon Sampling Theorem.

Digital to analogue conversion

Digital information is often put into analogue form by means of the 'sample and hold' technique. In this 2-stage technique, samples of the signal are first taken at regular intervals from computer memory, a disk or directly from the signal itself. The value of each sample is then held until the next one is read. To make matters more precise, let s denote the time between successive instants at which the samples are taken. Then the value $f(sk)$ of the signal f at the instant sk is held until the next sampling point $(s+1)k$ after which the next value $f((s+1)k)$ is held (see Figure 3). Mathematically this signal $b(t)$ say can be written

$$b(t) = \sum_{k=-\infty}^{\infty} f(sk)\chi_{[0,s)}(t-sk) = \sum_{k=-\infty}^{\infty} f(sk)\,\chi_{[sk,\,s(k+1))}(t), \tag{4}$$

where for any set A, $\chi_A(t)=1$ when $t\in A$ and 0 otherwise (χ_A is the characteristic or indicator function of the set A). This function is sometimes called a 'boxcar' signal (because of a resemblance to the profile of a US freight train) or sometimes a 'staircase' signal. Providing the sampling interval s is less than one half the reciprocal of the maximum frequency (as required in the WKS theorem), the boxcar signal has the same energy as the original signal.

The boxcar signal (4) is an approximation to the idealized representation in (3) of function $f(t)$ as the output of filtering a stream of weighted impulses. Actually, the boxcar signal is itself also only an idealization since it cannot be realized exactly, partly because the changes in the values of the samples will not be instantaneous and partly because of the 'quantization' errors already mentioned.

In the second stage, the boxcar signal formed from the discrete values of the digital information is smoothed by filtering out the high frequencies. When an analogue signal is being reconstructed from a boxcar signal in this way, frequencies above the maximum frequency of the original signal are filtered out. Energy is lost in the process and it turns out that more energy is lost in the high frequencies than the low, causing the reconstructed signal to be distorted. The difference between this signal and the original depends on the size of the sampling interval s. The smaller s the more accurate the reconstruction. In fact when s is small ($s<1/2W$) the error is proportional to about s^2, so that doubling the sampling rate (or halving the time interval) approximately quadruples the accuracy. Since s must be less than $1/2W$, the Sampling Theorem gives an indication of just how small s needs to be to ensure good reconstruction from the signal. In addition, the sampling rate of $2W$ (the Nyquist rate) given by the Sampling Theorem prevents aliasing, discussed below.

In practice sampling is carried out at a higher rate than the Nyquist rate. For example the sampling rate for compact discs is about 44 kHz (corresponding to the sampling interval s being about 1/44000th of a second). This is a little above the Nyquist rate for audio reconstruction since the upper limit of frequencies for conventional musical instruments is about 20 kHz.

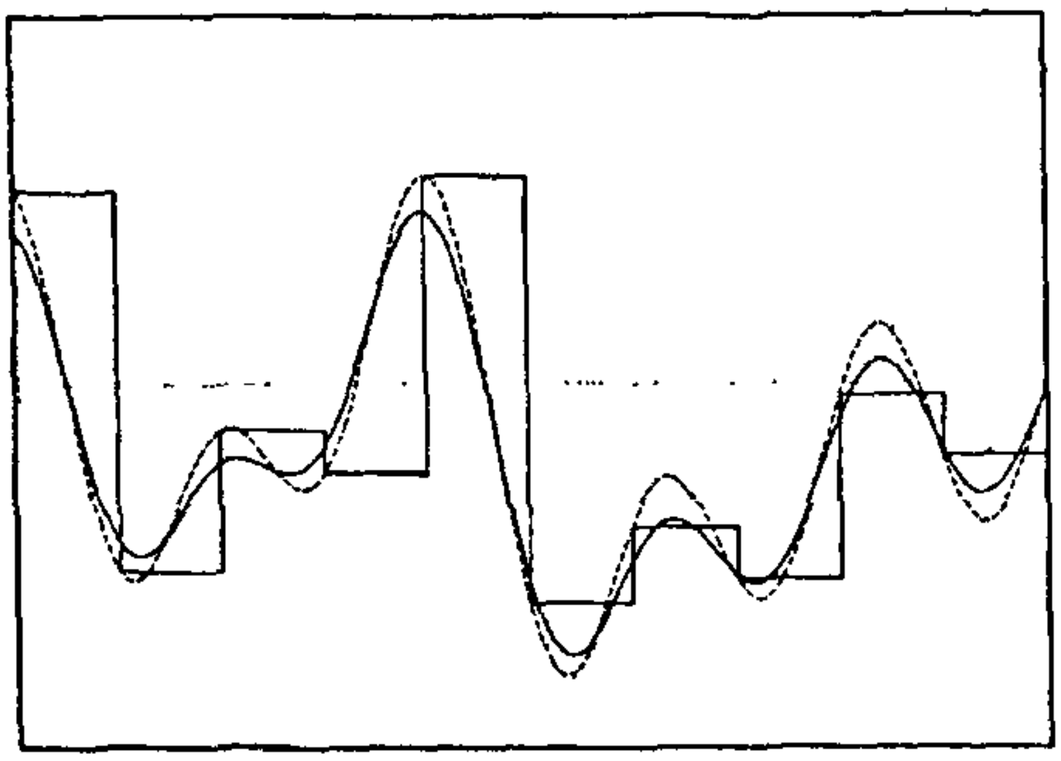

Figure 3. The original signal is shown by the dashed curve, the boxcar signal is formed by the 'sample and hold' technique. The heavy solid curve is the output after filtering the boxcar signal. The differences between the original and reconstructed curves are greatest near turning points.

Oversampling

Although ideally the Sampling Theorem implies that sampling above the Nyquist rate is unnecessary, the approximate reconstruction used in practice means that higher rates are often used. In fact sampling at a rate considerably higher than the Nyquist rate or *over-sampling* can pay dividends in several ways. It is used in some high fidelity compact disc players to avoid aliasing problems (see next section) associated with 'real' filters which do not cut off higher frequencies perfectly (refs. 26, 27, ch. 2). Three or seven additional samples are interpolated between successive samples being read off the disc, thus increasing the sampling rate 4 or 8-fold. This not only reduces errors of the kind just described but also improves other aspects of filter performance and allows tracking errors to be corrected.

Undersampling and aliasing

If a signal is sampled at less than the Nyquist rate or undersampled, then the resulting spectrum repetition in the frequency domain consists of translates of the spectrum which now overlap. As a result, high frequencies can appear as low frequencies. This effect is called 'aliasing' and can often be seen on films when rapidly moving wagon wheels appear to have slowly

revolving spokes. Stroboscopic lighting can actually exploit aliasing by apparently freezing rapid but repetitive motion. Because of the high frequencies masquerading as low frequencies, aliasing causes distortion in the reconstructed signal. Using the Nyquist rate (which is needed for the Sampling Theorem) guarantees that undersampling and hence aliasing will not occur.

Signal types

The bandlimited signals considered here fall into different types. Signals, such as speech, which have a spectrum consisting of a single band about the origin, with no gaps, are called *lowpass*. For purposes of transmission, radio and television frequencies are confined to a band centred about a high-frequency carrier wave. This type of signal is called *bandpass*. Signals which have spectra made up of several distinct bands such as that in Figure 4 are called *multiband* and are of increasing importance, for example they have been used in speech compression systems[28].

It is evident that their maximum frequency can be much greater than the total length of the bands making up the spectrum, whereas for a lowpass signal with maximum frequency W, the Nyquist rate of $2W$ is precisely the length $2W$ of the interval $(-W, W)$ of frequencies in the spectrum. Now it is this total length (or *measure*) of the frequency bands which is really fundamental in communication theory. Indeed in a remarkable paper Landau[29] proved that a signal could not be reconstructed stably from samples taken at a rate less than the measure of the support of the spectrum (length of the spectral bands) ('stably' means that small errors in the samples result in a signal close to the original). Of course a multiband signal can always be reconstructed from samples taken at the Nyquist rate of $2W$. This can be inefficient but there is an extension of the Sampling Theorem which can take advantage of the gaps in the spectrum to give lower sampling rates at the cost of a more complicated reconstruction[30].

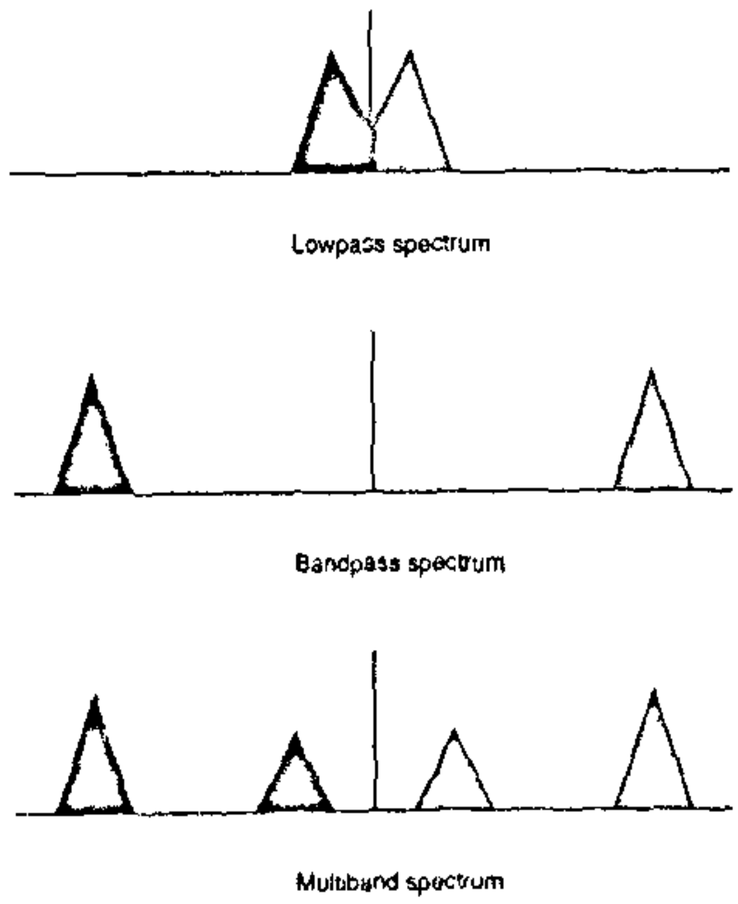

Figure 4. Different types of spectra.

It is a common practice to use higher, sometimes much higher sampling rates than the Nyquist rate, in order to avoid aliasing and to reduce attenuation caused by filtering. An awkward feature of multiband signals is that apparently reasonable choices of the sampling rate can lead to aliasing[31]. This means that for multiband signals care should be exercised in applying standard techniques appropriate to lowpass signals. For example, simply increasing the sampling rate will not necessarily prevent aliasing. The complete picture of which sampling rates will prevent aliasing and which won't depend on the spectral structure in a quite complicated way. But for multiband signals of a somewhat special type, with equally wide bands centred at the harmonics of a carrier frequency, the distribution of available sampling rates can be determined. Figure 5 shows sampling rate plotted against band separation with the available rates shown hatched; the rates in the 'bad' regions in between cause aliasing. It can be seen that hitting upon a reasonably small 'good' sampling rate lying in a hatched sector by chance is not very likely and so it is important to know the spectral structure.

Incidentally the figure is reminiscent of sets which arise in dynamical systems (such as the solar system) and which are associated with instability or chaos. There is a good reason for this. A sampling rate in a 'bad' region corresponds to a point which is 'close' to a rational with a small denominator. Integrally related quantities can cause resonance in a physical system, so that points close to rationals in this sense are 'near-

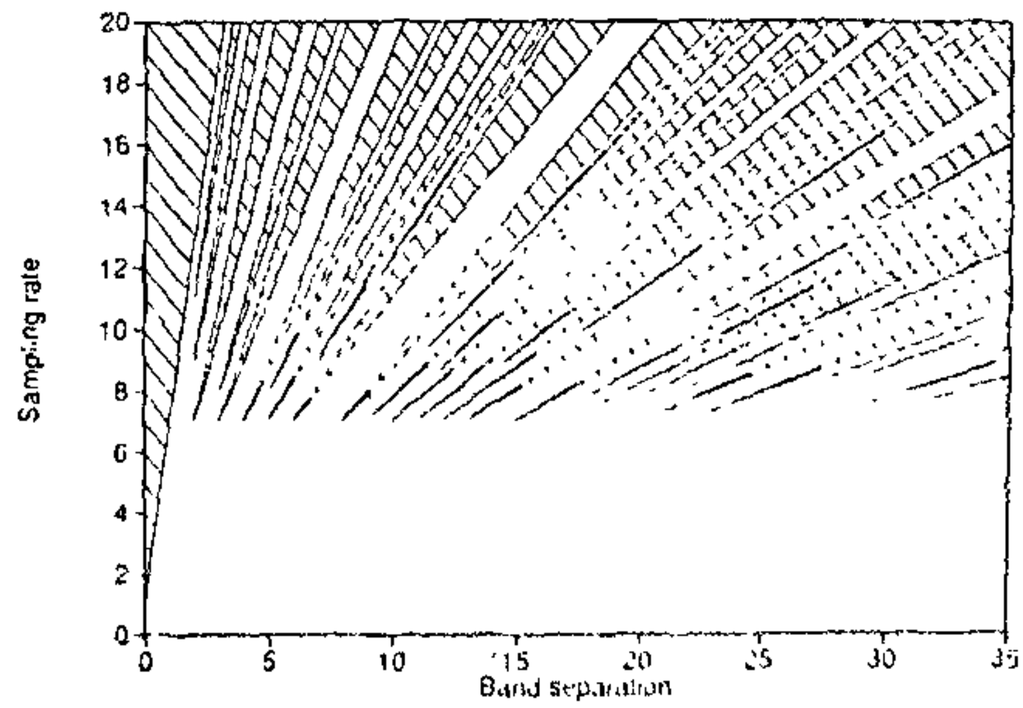

Figure 5. The hatched sectors correspond to sampling rates which avoid aliasing for signals with 3 equally wide, equally spaced frequency bands.

resonant' and can be associated with instability (e.g. vibration in a mechanical system). The 'closeness' of points to being rational is studied in Diophantine approximation, a branch of number theory. Thus as well as close connections with analysis, the Sampling Theorem has links with number theory and dynamical systems.

Conclusion

The Sampling Theorem was the key for extending results for discrete to analogue signals. It also provides appropriate sampling rates or channel widths for techniques used throughout signal processing. Moreover, although the Sampling Theorem cannot be implemented in exact mathematical form in electronic systems, there is a very simple and effective practical approximation which is very accurate for lowpass signals. Thus from both a theoretical and practical standpoint, the Sampling Theorem is a cornerstone of communication theory and engineering. As a bonus, it also lies at the heart of much mathematics.

ACKNOWLEDGEMENTS. It is a pleasure to acknowledge Professor Sivaraj Ramaseshan, who suggested this article, and Michael Beaty and Rowland Higgins who made many helpful comments on earlier drafts. Michael Beaty also prepared the figures and helped with preparing the text.

1. Shannon, C. E., *Bell Syst. Tech. J.*, 1948, **27**, 379–423 and 623–656.
2. Shannon, C. E., *Proc. IRE* 1949, **37**, 10–21.
3. Horgan, J., *IEEE Spectrum*, 1992, **29**, 72–75.
4. Khinchin, A. I., *Mathematical Foundations of Information Theory*, translated by R. A. Silverman and M. D. Friedman, Dover, New York, 1957.
5. Shannon, C. E. and Weaver, W., *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, 1949.
6. Blahut R. E., *Principles and Practice of Information Theory*, Addison-Wesley, Massachusetts, 1987.
7. Slepian, D., *Proc IEEE*, 1976, **64**, 292–300.
8. Kotel'nikov, V. A., Material for the First All Union Conference on Questions of Communication, in Russian, Izd. Red. Upr. Svyazi RKKA, Moscow, 1933 (Russian).
9. Raabe, H., *Elek. Nachrichtentechnik*, 1953, **6**, 213–228.
10. Weston, J. D., *Philos. Mag.*, 1949, **40**, 449–453.
11. Someya, I., *Waveform Transmission*, Shyukyoo, Tokyo, 1949, (Japanese).
12. Whittaker, J. M., *Cambridge Tracts in Mathematics and Mathematical Physics No. 33*, Cambridge University Press, Cambridge, 1935.
13. Landau, H. J., *Fourier Techniques and Applications* (ed. John F. Price), Plenum, New York, 1985, pp. 201–220.
14. Slepian, D., *SIAM Rev.*, 1983, **25**, 379–392.
15. Dym, H. and McKean H. P., *Fourier Series and Integrals*, Academic Press, New York, 1972.
16. Butzer, P. L., Hauss, M. and Stens, R. L., *Mathematical Sciences, Past and Present, 300 years of Mathematische Gesellschafts in Hamburg*, Mitteilungen Math. Ges. Hamburg, 1990.
17. Jerri, A. J., *Proc. IEEE*, 1977, **65**, 1565–1596.
18. Sayre, D., *Acta Crystallogr.*, 1952, **5**, 834.
19. Vaaler, J. D., *Bull. Am. Math. Soc.*, 1985, **12**, 183–216.
20. Balakrishnan, A. V., *IRE Trans. Inf. Th. IT-3*, 1957, 143–146.
21. Lloyd, S. P., *Proc. Am. Math. Soc.*, 1959, **92**, 1–12.
22. Higgins, J. R., *Bull. Am. Math. Soc.*, 1985, **12**, 45–89.
23. Whittaker, E. T., *Proc. R. Soc. Edinburgh*, 1915, **35**, 181–194.
24. Hardy, G. H., *Proc. Cambridge Philos. Soc.*, 1941, **37**, 331–348.
25. Butzer, P. L., *J. Math. Res. Exposition*, 1983, **3**, 185–212.
26. Goedhart, D., van de Plassche, R. J. and Stikvoort, E. F., *Philips Tech. Rev.*, 1982, **40**, 174–179.
27. Watkinson, J., *The Art of Digital Audio*, Focal Press, London, 1988.
28. Darnell, M., Dodson, M. M., Honary, B. and He, W., *Proc. Sixth International Conf. on System Engineering* (ed. Bellamy, N. W.), Coventry Polytechnic ICSE, 1988, pp. 109–119.
29. Landau, H. J., *Proc. IEEE*, 1967, **55**, 1701–1706.
30. Dodson, M. M. and Silva, A. M., *Proc. R. Ir. Acad.*, 1985, **85A**, 81–108.
31. Beaty, M. G. and Dodson, M. M., *SIAM J. App. Math.*, 1993, **53** (to appear).