The Establishment of Sampling as a Scientific Principle—A Striking Case of Multiple Discovery

Paulo J. S. G. Ferreira and Rowland Higgins

During the period 1928–1949, several engineers contributed to the establishment of a sampling principle. They did this virtually independently of each other in the context of communications theory and practice.

Five names stand out as being the main players in this drama: H. Nyquist, who laid the foundations for the minimal sampling rate; V. A. Kotel′nikov, C. E. Shannon, and I. Someya, whose treatment of the sampling principle was mathematical; and H. Raabe, who derived the minimal sampling rate and built hardware to reconstruct signals from samples taken at that rate. Their work is described below.

As Shannon recognized, the mathematical setting already existed. In fact, it was part of a tradition that can now be traced back to Cauchy, but its significance for application was simply not realized until Shannon's time.

Our notation is described in context, or else completely standard. For background and references to original work, see [1] and [3].

The Sampling Principle

I. Signal functions having finite energy and frequency content confined to a bounded set (e.g., $[-\pi w, \pi w]$, $w > 0$, the condition of *bandlimitation*) are uniquely determined by countably many of their values, or samples, taken at a fast enough rate.

II. Such a function can be represented, usually in the form of a series, in terms of these samples.

The Sampling Principle asserts that *all the energy, indeed all the information, in this type of function is contained in only countably many samples.* Seen in this way, the principle is one of data compression and is basic in modern digital communications.

The minimum sampling rate is w samples/second, twice the highest frequency component (measured in radians/second). The rate is usually called the Nyquist sampling rate.

The Sampling Principle has found widespread applications in modern science and technology, where it has been greatly extended and generalized.

When an idea comes to fruition at the hands of two or more people independently and at about the same time, historians call it *multiple discovery*, or *multiple invention.* Here, *discovery* seems the more appropriate choice. The Sampling Principle furnishes an example; we shall see that it grew out of the need to respond to the limitations set by contemporary technology, limitations which had to be understood and overcome, and that by the early twentieth century motives such as these were being felt worldwide. This may help to answer the question: *Why was the Sampling*

Paulo J. S. G. Ferreira is professor of telecommunications at Universidade de Aveiro, Portugal. His email address is `pjf@det.ua.pt`.

Rowland Higgins is emeritus professor at Anglia Ruskin University, Cambridge, England. His email address is `rhiggins11@gmail.com`.

Principle the subject of multiple discovery, spread over three continents?

Before going on to explore this question, we shall give a proof of the Sampling Principle in the language of today. This proof is a prototype of the elegant treatment, in terms of Hilbert spaces and their bases, that can be given to many general sampling theorems.

We have the standard notion of Hilbert space as a complete inner product space, the inner product being denoted here by $\langle \cdot, \cdot \rangle$ and the norm derived from the inner product denoted by $\| \cdot \|$.

A sequence (φ_n) in a Hilbert space H is an *orthonormal basis for* H if it satisfies both the orthonormality property

$$\langle \varphi_m, \varphi_n \rangle = \delta_{mn}, \quad m, n \in \mathbb{Z},$$

and the property that whenever $f \in H$, the representation

$$f = \sum_{n \in \mathbb{Z}} c_n \varphi_n$$

holds, with convergence in norm and with unique coefficients $c_n = \langle f, \varphi_n \rangle$.

To give the Sampling Principle a mathematical description, two Hilbert spaces are involved. One is $L^2(-\pi w, \pi w)$ (by a slight abuse of notation this symbol will mean those members of $L^2(\mathbb{R})$ that are null outside $[-\pi w, \pi w]$); the other is the Paley-Wiener space, denoted by PW, of continuous and square integrable functions on $\mathbb{R}$ whose Fourier transform is null outside $[-\pi w, \pi w]$, the norm being that of $L^2(\mathbb{R})$. Here the Fourier transform is defined by

$$(\mathcal{F} f)(\omega) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t) e^{-i\omega t} dt,$$

the integral taken in the L^2-sense.

It follows from the Plancherel theory of the L^2 Fourier transform that we can understand $f \in PW$ to be of the form

$$f(t) = \int_{-\pi w}^{\pi w} \varphi(\omega) e^{i\omega t} \, d\omega \tag{1}$$

for some $\varphi \in L^2(-\pi w, \pi w)$, and furthermore that the two Hilbert spaces are isometrically isomorphic under the transformation $\mathcal{F}$. In particular, an orthonormal basis for one space goes over into an orthonormal basis for the other.

Consider the orthonormal system

$$(2\pi w)^{-1/2} e^{-i\omega n/w}, \quad |\omega| \leq \pi w, \quad (n \in \mathbb{Z}), \tag{2}$$

the standard trigonometrical basis for $L^2(-\pi w, \pi w)$.

First, since a basis is, in particular, a complete set, we have

$$\int_{-\pi w}^{\pi w} \varphi(\omega) e^{i\omega n/w} \, d\omega = 0 \text{ for every } n$$

implies that φ is the null element of $L^2(-\pi w, \pi w)$. By (1) this means that $f(n/w) = 0$ for every n implies that φ is null. But if φ is null, so is f, and we have shown that $\{n/w\}$ *is a set of uniqueness for* PW. This accounts for part I.

Second, for Paley-Wiener functions, the classical sampling series is:

$$f(t) = \sum_{n \in \mathbb{Z}} f\left(\frac{n}{w}\right) \frac{\sin \pi(wt - n)}{\pi(wt - n)}. \tag{3}$$

To prove this we find, after a little calculation, that the (inverse) Fourier transform of the basis elements (2) (remembering that they are null outside $[-\pi w, \pi w]$) are proportional to the expansion functions in (3), and a little more calculation shows that the coefficients are indeed samples of f.

Thus the series (3) is an orthonormal expansion for $f \in PW$, and the corresponding Parseval relation is clearly

$$\|f\|_{L^2}^2 = \frac{1}{w} \sum_{n \in \mathbb{Z}} \left| f\left(\frac{n}{w}\right) \right|^2. \tag{4}$$

Convergence in (3) is in norm, but pointwise and uniform convergence can be obtained because the Cauchy-Schwarz inequality applies to (3), thanks to (4). This accounts for part II.

The Challenges of Communications Engineering

The answer to our question about multiple discovery must surely lie in the development of communications technology and theory during the interwar period.

Bandwidth limitation and its effect on the communication rate had been felt for the first time when telegraphy over submarine cables was attempted. The transmission speed was found to be severely limited because the cable acted like a capacitor, which had to charge and discharge before the signal could be received at the far end. This was experimentally shown by Faraday, in 1854. The delay reduced the transmission speed dramatically: the ninety-word message sent in 1858 across the Atlantic from Queen Victoria to President Buchanan took sixty-seven minutes to transmit. Rates of one or two words per minute were common.

Initially, this limitation was not fully understood. The Atlantic cable failed after only a few weeks, damaged by the high voltages used. Edward Orange Whitehouse, chief electrician with the Atlantic Telegraph Company, who had been convinced that only high voltages could deliver information across the Atlantic, was dismissed.

Kelvin's law of squares shed some light on the problem. It states that the maximum operating speed is proportional to $1/(RC\ell^2)$, where R and C are the cable's resistance and shunt capacitance per unit length and ℓ is its length. Since a simple RC circuit with resistance $R\ell$ and shunt capacitance $C\ell$ has bandwidth $\omega_0 = 1/(RC\ell^2)$, Kelvin's law states that *operating speed is proportional to bandwidth.*

The underlying theory was developed in 1854 in a correspondence between Stokes and Kelvin [4]. Assuming no inductance, Ohm's law is

$$-\frac{\partial V}{\partial x} = Ri,$$

where, as usual, V denotes the potential and i the current. A segment of unit length of cable at x accumulates charge at the rate $-\partial i/\partial x$. The potential therefore increases at the rate $-(1/C)\partial i/\partial x$, and so

$$C\frac{\partial V}{\partial t} = -\frac{\partial i}{\partial x}.$$

Eliminating i between these two equations, Kelvin obtained

$$RC\frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial x^2}, \tag{5}$$

which is similar to Fourier's equation for the propagation of heat. Kelvin knew Fourier's work very well and immediately recognized that it was "perfectly adapted" to the problem of the submarine cable. We outline two of the contributions found in the Kelvin–Stokes correspondence. First, the elementary solution of (5) is

$$V(x,t) = e^{-(RCn)^{1/2}x}\sin[2nt - (RCn)^{1/2}x]$$

and shows that harmonic terms of different frequencies are propagated at different velocities. The consequence is that no definite velocity of transmission is to be expected for more general signals, namely, linear combinations of such harmonic terms.

The second result, and the most important for us, is Kelvin's law of squares. To obtain it, he solved the diffusion equation for a unit step

$$V(0,t) = \begin{cases} 1, & t \geq 0, \\ 0, & t < 0, \end{cases}$$

and then computed the electrical current using

$$-\frac{\partial V}{\partial x} = Ri.$$

Kelvin determined that the current would reach a maximum after a certain time t_0, which he found by setting the derivative of the current to zero.

Kelvin's conclusion also follows from dimensional analysis: t_0 can depend only on ℓ, the length of the cable, and the product RC, the only parameter in the diffusion equation. The units into which the product RC can be expressed are

$$\frac{\text{unit of time}}{(\text{unit of length})^2}.$$

It follows that $RC\ell^2/t_0$ is invariant with respect to changes in the units of length and time. This means that it is a constant and so t_0 must be proportional to $RC\ell^2$, the law of squares.

Kelvin would apply Fourier's theory again in 1862, in a famous and controversial paper in which he used the equation of heat propagation to estimate the age of the Earth. It was in that paper that Kelvin referred to Fourier's work as a "great mathematical poem".

Heaviside, who considered a more complex heat propagation model and showed that the Earth could be much older than Kelvin had predicted, also improved Kelvin's telegraphy line model. He completed the model in 1887 by introducing series inductance. This decisive contribution allowed him to show that by carefully adding inductance to a cable its bandwidth could be increased. In fact, the attenuation of the cable could theoretically be made constant, that is, independent of the frequency.

Transmission speed was very important in telegraphy and the need for improvements globally felt. By 1896 there were about 160,000 nautical miles of cable, laid at a cost of $1,200 per mile, spread out all over the world, with London at the center. Due to theoretical progress and better instrumentation, the line between New York and the Azores Islands was being operated at four hundred words per minute in 1924.

Hartley and Nyquist made Kelvin's law precise. Between 1924 and 1928 they focused on abstract models of the channel rather than the physical properties of the cable. As a result, they captured in a precise way the interplay between transmission speed and bandwidth.

Progress in wireless telegraphy was also being made. Marconi's first transmissions across the Atlantic date from 1901. In Germany, Braun perfected the technology (and shared the 1909 Nobel Prize with Marconi). The discovery of the vacuum tube brought enormous progress, and a new form of bandwidth limitation was soon found: "the crowding of the ether", as Kotel′nikov put it. The wireless transmission of a signal required a certain bandwidth. The natural question was: How much bandwidth should be allocated to a certain signal? Conversely, given a certain bandwidth, how much information can be packed into it?

Telegraphy had exposed the effect of bandwidth limitations on the speed of transmission, leading to the results of Nyquist and Hartley. Telephony was exposing other forms of bandwidth limitation and raising new challenges.

Time-domain multiplexing had been used in telegraphy in order to allow the cables to transmit more than one signal simultaneously. The more complex multiplexing problem for telephony led Raabe in Germany to discover the minimum sampling rate for a given signal bandwidth.

Wireless transmission and frequency-domain multiplexing and the "crowding of the ether" pressed the engineers for precise answers regarding transmission rates, bandwidth, and the effect of noise. Answers appeared in Russia, the United States, and Japan by Kotel′nikov, Shannon, and Someya, respectively.

Shannon's work not only marks the birth of information theory, but it also represents a bridge between developments that can be traced back to telegraphy—Nyquist, Hartley, Kelvin—and mathematics—Fourier analysis and interpolation.

H. Nyquist

By 1928 Nyquist had identified the fundamental parameters that determine the rate at which information can be transmitted in a telegraph.

Nyquist observed that in telegraphy, time is divided into equal units. In each unit, the transmitted information identifies one symbol among m. He showed that the rate at which information can be transmitted increases linearly with both the number of symbols per second and the number of bits per symbol, $\log_2 m$. In 1928 Hartley reached similar conclusions independently, referring to the "considerable historical importance" of Kelvin's RC law.

Nyquist also clarified the connection between the transmission speed in telegraphy and bandwidth. This part of his work is closer in spirit to the sampling principle and led Shannon to coin the expression "Nyquist interval". The reciprocal of the Nyquist interval became known as the "Nyquist rate".

Nyquist considered the effect of transmitting periodic signals $f(t)$ and $f(2t)$ through two channels and realized that the channel used to transmit $f(2t)$ would have to deal with frequencies twice as large. He concluded that "frequency band is directly proportional to speed" and determined the proportionality constant by means of an argument involving Fourier series.

Nyquist's discoveries were motivated by concrete practical problems, but his focusing on the essential, abstract characteristics of telegraphy led him to general conclusions of lasting significance.

H. Raabe

By 1939 Raabe had built a multiplexing system for telephony, that is, a system to simultaneously transmit several signals over the same transmission line and recover each of them at the receiving end. The system worked by sampling each input signal in turn, at a certain fixed rate. The question that Raabe had to address was: At what rate should each signal be sampled?

In Raabe's system, each signal is multiplied by a square wave $s(t)$ that determines the sampling rate. He assumes that the input signal is periodic and that it can be expanded in a Fourier series. Multiplication of its components by the Fourier series of the square wave led Raabe to the answer: "distortionless transmission" is possible if the sampling frequency is at least twice the highest signal frequency.

To reach this conclusion, Raabe starts by writing the Fourier series of the square wave $s(t)$, translated to become an even function, as

$$s(t) = c + \sum_{n=1}^{\infty} \alpha_n \cos(n\omega_1 t), \tag{6}$$

where c is a nonzero constant—essentially, the average value of $s(t)$. The multiplexed signal is given by $r(t) = s(t)f(t)$. The question is of course whether $f(t)$ can be recovered from $r(t)$. To answer this question, Raabe expands $f(t)$ in a Fourier series. Instead of multiplying this Fourier series by that of $s(t)$, Raabe invokes superposition and investigates the multiplication of a single term of this Fourier series (one sinusoid, therefore) by that of $s(t)$. If this sinusoid is written as $\cos(m\omega_1 t)$, where m is a real number, the product becomes

$$\begin{aligned} r(t) &= c\cos(m\omega_1 t) \\ &\quad + \frac{1}{2}\sum_{n=1}^{\infty} \alpha_n\{\cos[(n-m)\omega_1 t] + \cos[(n+m)\omega_1 t]\}. \end{aligned}$$

Of the frequencies involved in this equation, $m\omega_1$ is due to the input signal, and the remaining frequencies, $(n \pm m)\omega_1$, can be regarded as "noise frequencies". Raabe observes that if $m\omega_1$ is known to fall below the smallest noise frequency, which is $(1-m)\omega_1$, there will be no problem in separating noise frequencies from signal frequencies. This leads him to the condition for lossless recovery: $m\omega_1 < (1-m)\omega_1$, that is, $m < 1/2$. In other words, the input frequency, $m\omega_1$, must be below one half the sampling frequency.

Raabe also found that band-pass signals (band-limited signals with no low-frequency terms) can be sampled at a lower rate, the first time this had been noted. These are important theoretical contributions in a work that has a remarkably strong practical character.

Raabe's finding that there is a minimum sampling frequency for low-pass and band-pass signals that, in theory, allows distortionless transmission is another instance of a theoretical discovery prompted by practical needs; in this case, the multiplexing problem.

V. A. Kotel′nikov, C. E. Shannon, and I. Someya

These three engineers introduced the sampling series into communications engineering independently of each other; Kotel′nikov in 1933, Shannon in 1949, and Someya also in 1949. Their proofs differ, of course, and none of them is strictly rigorous, but all are directly appealing to the intuition. Some minor changes in the original notations have been made.

Kotel′nikov's hypotheses are that $f \in L^1(\mathbb{R})$, f satisfies Dirichlet's conditions and is band-limited to $[-\pi w, \pi w]$.

His proof is based on a Fourier inversion principle for such functions, quoted from the classical literature, to the effect that f satisfies (1) where

$$\varphi(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt. \tag{7}$$

Next he writes down the Fourier series for φ, the coefficients being $(f(n/w))$ from (1). This Fourier series is now substituted into (1), and (3) results.

An interesting feature of the proof, not found in the other two, is that Kotel′nikov recognizes the presence of a converse to the sampling theorem, that is, that if a function f is represented by a series of the form (3), with $(f(n/w))$ replaced by a numerical sequence (D_n), then it must be band-limited. He argues that, since every term of the series is band-limited to $[-\pi w, \pi w]$ (by a special Fourier transform), the same must be true of its sum.

Shannon works with what we have called Paley–Wiener functions. An interesting feature of Shannon's argument, not found in the other two, is that he shows $(1/w)\mathbb{Z}$ to be a set of uniqueness for PW (this is Part I of the Sampling Principle). He does this by following through a chain of unique determinations: f is uniquely determined by its Fourier transform, which in turn is uniquely determined by its Fourier coefficients, which in turn are uniquely determined by samples of f at (scaled) integer time points. That is, f is uniquely determined by its samples.

As for (3), Shannon argues that the sum is band-limited, just as Kotel′nikov did. This sum coincides with f at the sample points (a simple calculation); therefore, by the uniqueness proved in the first part, the sum is f.

Someya works with functions that he designates as being merely "band-limited", nothing more. His proof is similar in outline to that of Shannon but differs in detail; in fact, it is obscure and unnecessarily lengthy, and it will not be feasible to give a complete account of it here (see [2] for an assessment of this proof). However, the important fact remains that Someya's contribution is a completely independent introduction of the sampling theorem in the engineering context.

Conclusion

We have asked a historical question, but history seldom provides us with clear-cut answers. However, the emergence of sampling in practice seems to be closely connected to the development of communications engineering. Bandwidth limitation was first felt in connection with submarine telegraphy and led to the results of Kelvin and Heaviside. By 1928 Hartley and Nyquist were taking a more general approach to telegraphy, linking transmission speed and bandwidth in a precise way.

As wireless telegraphy and telephony began to develop, new forms of bandwidth limitation were found. The two simplest ways of sharing a channel are time-division and frequency-division multiplexing. The former stimulated the work of Raabe. The latter, in which different signals are assigned different band-regions, raises the question of how much bandwidth needs to be allocated to a signal. We have seen that Kotel′nikov, Someya, and Shannon addressed the problem. The work of Shannon, in particular, established a bridge between developments that had their origin in telegraphy, multiplexing, mathematics, and signal analysis.

The recent growth in bandwidth usage due to the widespread use of mobile devices is raising new challenges. The "crowding of the ether" is a problem as pressing today as it was in Kotel′nikov's time. Bandwidth remains precious: half of the 108 MHz of prime spectrum freed thanks to the recent shift to digital television in the United States was auctioned by the U.S. Treasury and sold for $19 billion. Telecommunications, as a source of problems of theoretical interest and practical importance, has not yet been exhausted.

References

1. P. L. Butzer, M. M. Dodson, P. J. S. G. Ferreira, J. R. Higgins, O. Lange, P. Seidler, and R. L. Stens, Multiplex signal transmission and the development of sampling techniques: The work of Herbert Raabe in contrast to that of Claude Shannon, *Applicable Analysis* **90** (2011), no. 3, 643–688.
2. P. J. S. G. Ferreira, *Someya's Proof of the Sampling Theorem—A Critique*, Tech. report, University of Aveiro, Portugal, July 2007.
3. J. R. Higgins, Five short stories about the cardinal series, *Bull. Amer. Math. Soc.* **12** (1985), 45–89.
4. W. Thomson, On the theory of the electric telegraph, *Proc. Royal Soc. London* **7** (1854–1855), 382–399, also reprinted in Lord Kelvin's *Mathematical and Physical Papers*, vol. II, p. 61.