Some Comments on Fourier Analysis, Uncertainty and Modeling

David Slepian

SIAM Review, Volume 25, Issue 3 (Jul., 1983), 379-393.

SIAM REVIEW
Vol. 25, No. 3, July 1983

0036-1445/83/2503-0005 $01.25/0

SOME COMMENTS ON FOURIER ANALYSIS, UNCERTAINTY AND MODELING*

DAVID SLEPIAN†

Abstract. Investigation of the problem of simultaneously concentrating a function and its Fourier transform has led to some interesting special functions that have widespread applications in engineering. They provide a means of proving a rigorous version of an engineering folk theorem called the $2WT$-Theorem. Many generalizations of these ideas seem to possess a similar elegant mathematical structure. A brief descriptive review is given of these developments.

Introduction. Let me begin by giving my thanks to Professor Parter and to the John von Neumann Lecture Committee for bestowing this honor on me. I was truly moved by this unexpected recognition.

I am going to use this occasion to tell you in detail about a problem in Fourier analysis that arose in a quite natural manner in a corner of electrical engineering known as Communication Theory. The problem was first attacked more than 20 years ago [1]–[3] jointly by me and two colleagues at Bell Labs—Henry Pollak and Henry Landau. It differed from other problems I have worked on in two fundamental ways. First, we solved it—completely, easily and quickly. Second, the answer was interesting—even elegant and beautiful. (Usually I struggle for months or years with a problem. If I do "solve" it, it is usually only in part, and the answer itself is rarely interesting. The interest generally lies in the fact that I have proved that this is the answer.) In the case of this problem, however, the answer had so much unexpected structure that we soon saw that we had solved many other problems as well. We had answered questions we had not meant to ask in optics, estimation and detection theory, quantum mechanics, laser modes—to name a few.

There was a lot of serendipity here, clearly. And then our solution, too, seemed to hinge on a lucky accident—namely that we found a second-order differential operator that commuted with an integral operator that was at the heart of the problem. Soon afterwards, a number of obvious generalizations of the original problem [4] yielded to the same techniques. They had answers with the same elegant structure, and again there was by good luck a commuting second-order differential operator.

We had scratched the ground a bit and had unknowingly uncovered the tip of a rich vein of ore. Off and on for the next 20 years I came back and mined a new piece of it [5]–[10]. Nor is it exhausted yet. In recent years Professor F. A. Grünbaum and his students have dug their shovels in and have unearthed interesting novel generalizations and ramifications of the original problem [11]–[15]. The mystery of this serendipity grows. Most of us feel that there is something deeper here than we currently understand—that there is a way of viewing these problems more abstractly that will explain their elegant solution in a more natural and profound way, so that these nice results will not appear so much as a lucky accident. But more of that later.

The role of Fourier analysis in electrical engineering. I would like to motivate my discussion of the problem by saying a few words about the use of Fourier analysis in electrical engineering. It is of paramount importance in that discipline.

*Received by the editors October 25, 1982, and in revised form December 6, 1982. This is a written and slightly expanded version of the John von Neumann Lecture given July 20, 1982 at the SIAM 30th Anniversary Meeting, Stanford University, Stanford, California.

†Bell Laboratories, Murray Hill, New Jersey 07974, retired.

Communication engineers are much concerned with the class of functions they call *signals*. These are real functions, $r(t)$, defined everywhere on a real line that they call *time*. They are square-integrable there,

$$\int_{-\infty}^{\infty} r^2(t)\,dt = E < \infty \tag{1}$$

and the quantity E is called the *energy* of the signal $r(t)$. It has an important physical interpretation. Typically, $r(t)$ will represent the voltage difference at time t between two points in an electrical circuit or device—say the voltage difference between the terminals of a microphone, for example. *Signal space* S, is the set of all signals, $r(t)$. It is, of course, just the space $L^2(-\infty, \infty)$ of the mathematician.

Signals possess Fourier transforms,

$$R(f) = \int_{-\infty}^{\infty} e^{-2\pi ift} r(t)\,dt, \tag{2}$$

which themselves are integrable in absolute square, and from Parseval's relation we have

$$\int_{-\infty}^{\infty} |R(f)|^2\,df = \int_{-\infty}^{\infty} |r(t)|^2\,dt = E. \tag{3}$$

The quantity $R(f)$ is known as the *amplitude spectrum* of the signal, and plays a key role in the engineer's analysis. He often works more with $R(f)$ than with the original signal. The inverse Fourier transform

$$r(t) = \int_{-\infty}^{\infty} R(f) e^{2\pi ift}\,df \tag{4}$$

allows him to think of the signal $r(t)$ as a sum of sinusoids of different frequencies. The *sinusoid* of frequency f, namely $e^{2\pi ift}$, has *amplitude* $|R(f)|$ and *phase* $\arg R(f)$.

Engineers build devices and networks that transform signals into new signals of some desired form. They think of their devices as operating on an input signal $r(t)$ to produce an output signal $s(t)$,

$$s(t) = M[r(t)]$$

where M is an operator. $s(t)$ is often called the *response* of the system M to the input $r(t)$. Fortunately, it is frequently the case that large portions of the structures of interest to the engineer can be well approximated by operators that are linear and time-invariant. If L is such an operator, this means that

$$s_i(t) = L[r_i(t)], \qquad i = 1, 2 \tag{5}$$

implies that

$$as_1(t) + bs_2(t) = L[ar_1(t) + br_2(t)] \tag{6}$$

and

$$s_1(t - T) = L[r_1(t - T)] \tag{7}$$

for all real T, all complex constants a and b, and all inputs $r_1(t)$, $r_2(t)$.

The response $s_f(t)$, of a linear time-invariant system to the sinusoidal input $e^{2\pi ift}$ is easily calculated. We have

$$s_f(t) = L[e^{2\pi ift}],$$

so that

$$s_f(t + T) = L[e^{2\pi if(t+T)}] = L[e^{2\pi ifT} e^{2\pi ift}] = e^{2\pi ifT} L[e^{2\pi ift}] = e^{2\pi ifT} s_f(t) \tag{8}$$

from (7) and from (6) with $b = 0$ and $a = e^{2\pi i f T}$. Set $t = 0$ in (8) to obtain $s_f(T) = e^{2\pi i f T} s_f(0)$. Since this must hold for all T, replacing T by t we get

$$s_f(t) = L[e^{2\pi i f t}] = Y_L(f) e^{2\pi i f t} \tag{9}$$

where I have written $Y_L(f) \equiv s_f(0)$. The response to $e^{2\pi i f t}$ is a sinusoid of the same frequency but with a complex amplitude $Y_L(f)$ characteristic of the operator in question. Thus we have the fundamental fact (9) that *all linear time-invariant operators have the complex exponentials* $e^{2\pi i f t}$, $-\infty < f < \infty$, *as eigenfunctions. Different linear time-invariant operators are characterized by their eigenvalues* $Y_L(f)$ (called the *transfer function* of L). This fact is the true genesis of the widespread applicability of Fourier analysis to electrical engineering.

Since every $r \in S$ can be expanded in the eigenfunctions of linear time-invariant operators in the sense of (4), the response of such an operator to any signal r is easily calculated from (9):

$$\begin{aligned} s(t) = L[r(t)] &= \int_{-\infty}^{\infty} R(f) L[e^{2\pi i f t}]\, df \\ &= \int_{-\infty}^{\infty} R(f) Y_L(f) e^{2\pi i f t}\, df. \end{aligned}$$

The amplitude spectrum, $S(f)$, of the response of L to input $r(t)$ is therefore

$$S(f) = Y_L(f) R(f), \tag{10}$$

and so the effect of a linear time-invariant operator is described in terms of the amplitude spectrum of signals by a simple multiplication by the transfer function $Y_L(f)$.

Bandwidth and uncertainty. Physical devices do not transmit sinusoids of arbitrarily high frequency without severe attenuation. The transfer functions $Y_L(f)$ used by engineers tend to zero with increasing f. It follows from (10) that the amplitude spectra of the responses of these systems to signals of finite energy also are negligibly small beyond some finite frequency.

Examination of the most natural classes of input signals shows that they too have amplitude spectra of finite support. For example, Fourier analysis of recorded male speech gives an amplitude spectrum that is zero for frequencies higher than 8000 hertz (= cycles/second). Conventional orchestral music has no frequencies higher than 20,000 hertz, while the output of a television camera (vidicon) has an amplitude spectrum vanishing for $|f| > 2 \times 10^6$ hertz.

Thus, both because of the frequency limiting nature of the devices he uses and because of the nature of the signals he is interested in transmitting, the communication engineer is soon led to consider the *space* B_W *of bandlimited signals*. These are the $r \in S$ whose amplitude spectra vanish for $|f| > W$. Each member of B_W can be written as a finite Fourier transform:

$$r(t) = \int_{-W}^{W} e^{2\pi i f t} R(f)\, df. \tag{11}$$

Here $W > 0$ is a parameter which I shall regard as fixed in our further discussion. Members of B_W are said to be *of bandwidth* W or are said to be *bandlimited to the band* $(-W, W)$. In an analogous manner, $r(t)$ is said to be *timelimited* if for some $T > 0$, $r(t)$ vanishes for all $|t| > T/2$.

It follows readily from (11) that bandlimited signals are extremely smooth. Indeed, one can allow t to be complex in (11) and it is easily seen that this extended $r(t)$ is an entire function of the complex variable t. It has no singularities in the finite t-plane, is infinitely differentiable everywhere, and has a Taylor series about every point with an infinite radius of convergence. From this it follows that a nontrivial bandlimited signal cannot vanish on any interval of the t-axis. For if it did, it and all its derivatives would be zero at some interior point of the interval, and a Taylor expansion would require it to be the trivial everywhere-zero signal. *The only signal that is both bandlimited and timelimited is the trivial always-zero signal.*

Now, in addition to transmitting smooth signals like speech or music, communication engineers would like to deal with short pulses such as the dots and dashes of telegraphy or the time markers of radar. As we have seen, such a signal of finite time-support cannot be bandlimited—it must contain sinusoids of arbitrarily large frequency. And bandlimited signals cannot have finite time-support—they must wander on forever. This dilemma has long plagued the communication engineer as he has sought signals that are somehow concentrated in both the time and frequency domains.

There are many ways of mathematically expressing this impossibility of simultaneous confinement of a signal and its amplitude spectrum. The most famous is the Heisenberg uncertainty principle of quantum mechanics. Its mathematical content is as follows. Let the signal $r(t)$ have unit energy, so that

$$\int_{-\infty}^{\infty} r^2(t)\,dt = \int_{-\infty}^{\infty} |R(f)|^2\,df = 1.$$

Both $r^2(t)$ and $|R(f)|^2$ can be regarded then as probability densities and their variances σ_r^2 and σ_R^2 can be regarded as measures of the extent to which the densities are concentrated. The Heisenberg uncertainty principle simply states that

$$\sigma_r^2\sigma_R^2 \geqq \frac{1}{(4\pi)^2}.$$

This result follows directly from the Schwarz inequality and the definition of variance

$$\sigma_r^2 \equiv \int_{-\infty}^{\infty} (t-t_0)^2 r(t)^2\,dt, \qquad t_0 \equiv \int_{-\infty}^{\infty} t r(t)^2\,dt. \tag{12}$$

The concentration problem. The measure (12) of concentration fits nicely into the concepts of quantum mechanics, but has little physical significance for the communication engineer. A more natural and meaningful measure of concentration of a signal for him is

$$\alpha^2(T) \equiv \frac{\int_{-T/2}^{T/2} r^2(t)\,dt}{\int_{-\infty}^{\infty} r^2(t)\,dt}, \tag{13}$$

i.e., the fraction of the signal's energy that lies in a given time slot. Similarly

$$\beta^2(W) \equiv \frac{\int_{-W}^{W} |R(f)|^2\,df}{\int_{-\infty}^{\infty} |R(f)|^2\,df} \tag{14}$$

is an appropriate measure of concentration of the amplitude spectrum of $r(t)$.

If $r(t)$ were indeed timelimited to $(-T/2, T/2)$, $\alpha^2(T)$ in (13) would have its largest value, namely unity. A nontrivial bandlimited signal cannot be so timelimited, however, and a very natural question is to determine how large $\alpha^2(T)$ can be for $r(t) \in B_W$. This, finally, is the problem Landau, Pollak and I considered these many years ago. To answer it, first express $r(t)$ in (13) in terms of its amplitude spectrum $R(f)$ via (11) and (3)

$$\alpha^2(T) = \frac{\int_{-T/2}^{T/2} dt \int_{-W}^{W} df'' e^{2\pi i f'' t} R(f'') \int_{-W}^{W} df' e^{-2\pi i f' t} R(f')^*}{\int_{-W}^{W} |R(f)|^2 \, df}$$

$$= \frac{\int_{-W}^{W} df' \int_{-W}^{W} df'' \frac{\sin \pi T(f' - f'')}{\pi(f' - f'')} R(f'') R(f')^*}{\int_{-W}^{W} df' R(f') R(f')^*}.$$

Here now $R(f)$ is an arbitrary function in $L^2(-W, W)$, and the problem of maximizing $\alpha^2(T)$ has been reduced to a classic form [16]. A maximizing R must satisfy the integral equation

$$\int_{-W}^{W} \frac{\sin \pi T(f' - f'')}{\pi(f' - f'')} R(f'') \, df'' = \alpha^2(T) R(f'), \qquad |f'| \leqq W, \tag{15}$$

a homogeneous Fredholm equation of the second kind. To investigate it further, we make the substitutions and definitions

$$f'' = Wy, \quad f' = Wx, \quad R(Wy) \equiv \psi(y), \quad \alpha^2(T) \equiv \lambda \tag{16}$$

$$c \equiv \pi WT \tag{17}$$

which reduce (15) to

$$\int_{-1}^{1} \frac{\sin c(x - y)}{\pi(x - y)} \psi(y) \, dy = \lambda \psi(x), \qquad |x| \leqq 1, \tag{18}$$

in which a single parameter, c, appears instead of W and T separately as in (15).

The prolate spheroidal wave functions. It is easy to show that the symmetric kernel $(\sin c(x - y))/\pi(x - y)$ is positive definite, so that from standard theory [16] we know that (18) has solutions in $L^2(-1, 1)$ only for a discrete set of real positive values of λ, say $\lambda_0 \geqq \lambda_1 \geqq \lambda_2 \geqq, \cdots$ and that as $n \to \infty$, $\lim \lambda_n = 0$. The corresponding solutions, or *eigenfunctions*, $\psi_0(x), \psi_1(x), \psi_2(x), \cdots$ can be chosen to be real and orthogonal on $(-1, 1)$. They are complete in L^2 there.

The variational problem that led to (18) only requires that equation to hold for $|x| \leqq 1$. With $\psi(y)$ on the left of (18) given for $|y| \leqq 1$, however, the left is well defined for all x. We use this to extend the range of definition of the ψ's and so define

$$\psi_n(x) \equiv \frac{1}{\lambda_n} \int_{-1}^{1} \frac{\sin c(x - y)}{\pi(x - y)} \psi_n(y) \, dy, \qquad |x| > 1.$$

The eigenfunctions $\psi_n(x)$ are now defined for all x and a simple calculation shows that they are orthogonal on $(-\infty, \infty)$ as well as on $(-1, 1)$ as already noted. We normalize them to have unit energy, so that

$$\int_{-\infty}^{\infty} \psi_n(x) \psi_m(x) \, dx = \delta_{mm}, \tag{19}$$

and it then follows that

$$\int_{-1}^{1} \psi_n(x) \psi_m(x) \, dx = \lambda_n \delta_{mn}. \tag{20}$$

This remarkable double orthogonality of the ψ_n is useful in many applications. Note that λ_n is the fraction of the energy of ψ_n that lies in the interval $(-1, 1)$.

It is difficult to obtain much detailed information about the ψ's directly from (18). Fortunately they are also solutions to a second-order differential equation eigenvalue problem,

$$\frac{d}{dx}(1 - x^2)\frac{d\psi}{dx} + (\chi - c^2x^2)\psi = 0, \tag{21}$$

and from this equation one can deduce much about them. Equation (21) had been much studied previously [17]–[18] since it arises on separating the 3-dimensional scalar wave equation in a prolate spheroidal coordinate system. It has solutions bounded everywhere only for discrete real positive values of the parameter χ, say $0 < \chi_0 < \chi_1 < \chi_2 < \cdots$. The corresponding solutions $\psi_0, \psi_1, \psi_2, \cdots$ are known as *prolate spheroidal wave functions* (pswf's). They are also solutions of (18) and, when indexed by increasing values of χ, they agree with the notation of indexing by decreasing values of λ as already indicated. That the solutions of (21) satisfy (18) can be deduced from the fact that the ψ's are complete in $L^2(-1, 1)$ and the fact that the differential operator

$$P_x \equiv \frac{d}{dx}(1 - x^2)\frac{d}{dx} - c^2x^2$$

commutes with the kernel of (18), i.e. that for all signals $r(y)$, and all real x,

$$P_x \int_{-1}^{1} \frac{\sin c(x - y)}{\pi(x - y)} r(y)\, dy = \int_{-1}^{1} \frac{\sin c(x - y)}{\pi(x - y)} P_y r(y)\, dy.$$

This is the seemingly lucky accident referred to in the Introduction.

From (18) and (21) one can deduce many properties of the $\psi_n(x)$ and λ_n which I now state with no indication of proof:

$$\begin{aligned} &\lambda_0 > \lambda_1 > \lambda_2 > \cdots, \\ &\psi_n(x) \text{ is even or odd with } n, \\ &\psi_n(x) \text{ has exactly } n \text{ zeros in } (-1, 1), \\ &\psi_n(x) \sim k_n \frac{\sin cx}{x}, \text{ as } x \to \infty, \end{aligned} \tag{22}$$

$$\int_{-1}^{1} e^{2\pi i x t}\psi_n(t)\, dt = \alpha_n \psi_n(2\pi x/c), \quad -\infty < x < \infty, \tag{23}$$

where k_n and α_n are independent of x. This last equation states the curious fact that the Fourier transform of $\psi_n(t)$ restricted to $|t| < 1$ has the same form as ψ_n except for a scale change. The equation can also be written

$$\psi_n(x) = \frac{1}{\alpha_n}\int_{-c/2\pi}^{c/2\pi} e^{2\pi i x t}\psi_n(2\pi t/c)\, dt, \tag{24}$$

which shows $\psi_n(x)$ to be bandlimited of bandwidth $c/2\pi$. The $\psi_n(x)$ are complete in $-\infty < x < \infty$ in the L^2 sense among signals $r(x)$ of bandwidth $c/2\pi$ or less, i.e. complete in $B_{c/2\pi}$.

It should be noted that both the $\psi_n(x)$ and λ_n depend on the parameter c, although I have suppressed this fact by my notation. Indeed, the dependence of these quantities on c is of considerable interest. Figure 1 shows $\psi_8(x)$ for $0 \leqq x \leqq 1.5$ for 5 different c values.

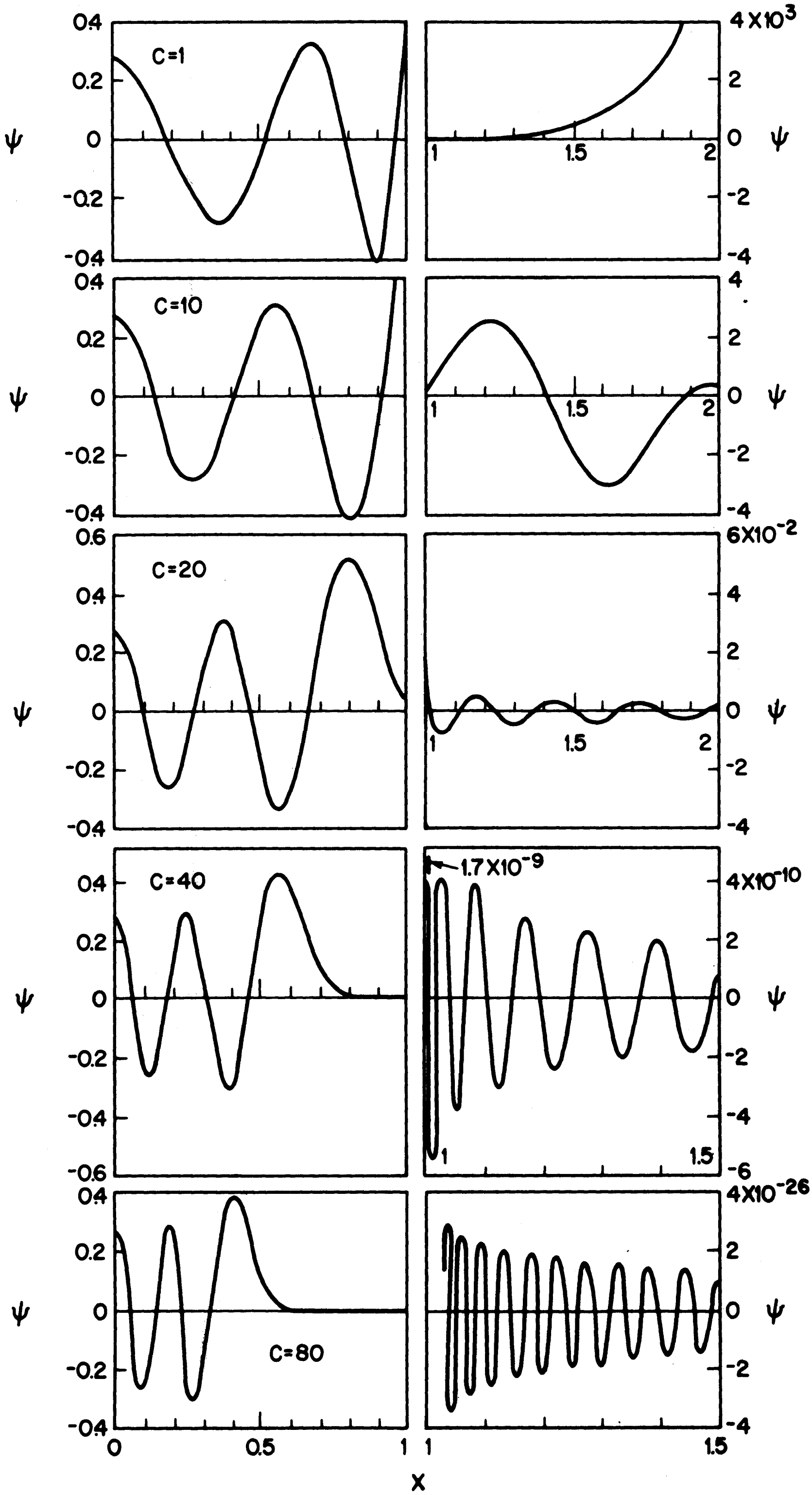

FIG. 1. $\psi_8(x)$ *for* $c = 1, 10, 20, 40, 80$. *Note changes in vertical scales for* $x > 1$ *for each* c *and in horizontal scales for* $x > 1$ *for* $c = 40, 80$.

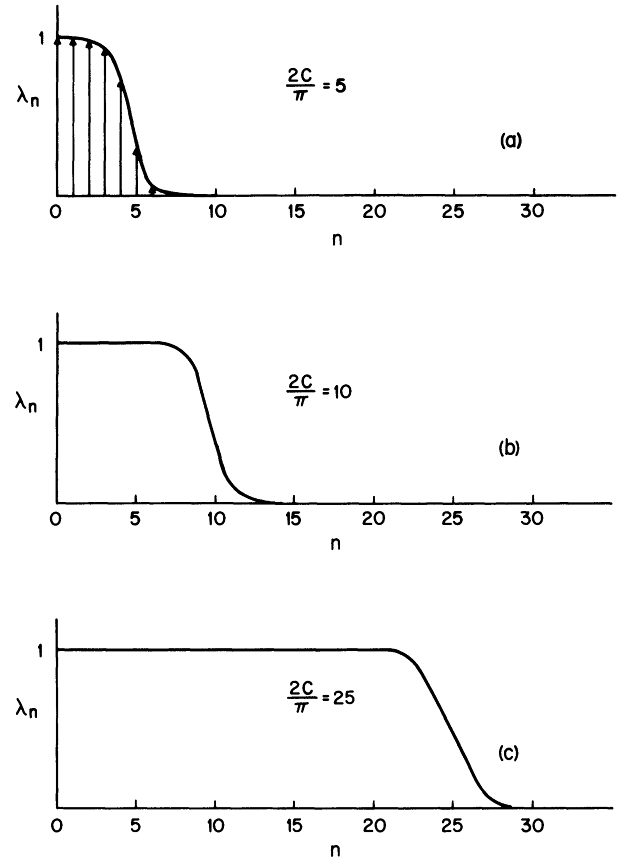

FIG. 2. *Dependence of* λ_n *on c.*

The behavior of $\psi_8(x)$ for x in (0, 1) is very different from its behavior for $x > 1$. Instead of following the convention (19) of giving ψ_8 unit energy, I have in Fig. 1 normalized it so that $\psi_8(0) = .27$ which permits observation of the detail within the interval $0 \leqq x < 1$. There are 4 zeros of $\psi(x)$ in this interval in accordance with (22). As c increases, the zeros move closer to the origin. Note well for $x > 1$ the change of scale with c in Fig. 1. For $c = 1$, $\psi_8(x)$ becomes enormous around $x = 2$ before finally falling away like $k(\sin x)/x$. When $c = 10$, ψ_8 rises only to a value approximately 2 before it decays. For the larger values of c shown, ψ_8 is minuscule for $x > 1$.

The foregoing has illustrated that for large c almost all the energy of ψ_8 lies in $(-1, 1)$; for very small c, almost all the energy of ψ_8 lies outside $(-1, 1)$. This dependence of concentration on c is made more transparent by noting how λ_n depends on c. Figure 2 shows the λ's for $c = 5\pi/2$, $10\pi/2$ and $25\pi/2$. The arrows on Fig. 2a give the values of $\lambda_0, \lambda_1, \cdots, \lambda_6$. I have drawn a smooth curve through the tops of these arrows. The analogous curve alone is shown for two other c values in Figs. 2b and 2c. It is seen from these curves that for $n \ll 2c/\pi$ most of the λ_n are close to unity while for $n \gg 2c/\pi$ most of the λ_n are nearly zero. When $n \approx 2c/\pi$, $\lambda_n \approx 1/2$. It is also seen from the figure that as c gets large the width of the n-interval in which λ_n falls from one to zero increases. It was established very early [3] that this interval grows like log c, and a few years later [6] I "showed" that

(25)
$$\lim_{c\to\infty} \lambda_n = \begin{cases} 0, & n = \left[(1+\eta)\dfrac{2c}{\pi}\right], \\ [1+e^{\pi b}]^{-1}, & n = \left[\dfrac{2c}{\pi} + \dfrac{b}{\pi}\log c\right], \\ 1, & n = \left[(1-\eta)\dfrac{2c}{\pi}\right], \end{cases}$$

where $\eta > 0$ is an arbitrarily small fixed number. Here $[x]$ = largest integer $\leqq x$. My demonstration was far from rigorous; asymptotic series were manipulated without regard for convergence, etc. But numerical work left little doubt about the validity of the formula. It was finally established rigorously by Landau and Widom in 1980 [19].

Concentrated bandlimited functions. The prolate spheroidal wave functions $\psi_n(x)$ just discussed provide a very useful set of bandlimited signals for the communications engineer. Let $W > 0$ and $T > 0$ be given. Define

$$r_n(t) \equiv \sqrt{\frac{2}{T}}\,\psi_n\left(\frac{2t}{T}\right), \qquad c \equiv \pi W T.$$

Then

(26)
$$r_n(t) \in B_W,$$
$$\int_{-T/2}^{T/2} r_n(t) r_m(t)\, dt = \lambda_n \delta_{mn},$$
$$\int_{-\infty}^{\infty} r_n(t) r_m(t)\, dt = \delta_{mn},$$
the $r_n(t)$ are complete in B_W for $-\infty < t < \infty$,

the $r_n(t)$ are complete in $L^2\left(-\frac{T}{2}, \frac{T}{2}\right)$,
$$\int_{-\infty}^{\infty} e^{2\pi i f t} r_n(t)\, dt = c_n \begin{cases} r_n\left(\dfrac{T}{2}\dfrac{f}{W}\right), & |f| \leqq W, \\ 0, & |f| \geqq W, \end{cases}$$
$$\int_{-T/2}^{T/2} e^{2\pi i f t} r_n(t)\, dt = d_n r_n\left(\frac{T}{2}\frac{f}{W}\right), \qquad -\infty < f < \infty,$$
$$\int_{-T/2}^{T/2} \frac{\sin 2\pi W(t-t')}{\pi(t-t')} r_n(t')\, dt' = \lambda_n r_n(t), \qquad -\infty < t < \infty,$$

Among signals in B_W, $r_0(t)$ is the most concentrated in $(-T/2, T/2)$ and its concentration is $\alpha^2(T) = \lambda_0$. Among signals in B_W that are orthogonal to $r_0(t)$, $r_1(t)$ is most concentrated; for it, $\alpha^2(T) = \lambda_1$. In general, $r_n(t)$, whose concentration is λ_n, is the most concentrated signal in B_W that is orthogonal to $r_0, r_1, \cdots, r_{n-1}$.

These special signals have many extremal properties and provide ready solutions to many natural problems. I mention only two here.

i) For $r(t) \in S$, what are the possible values of $\alpha^2(T)$ and $\beta^2(T)$ in (13) and (14)? They cannot both be unity, for example, for this would imply a nontrivial signal that was

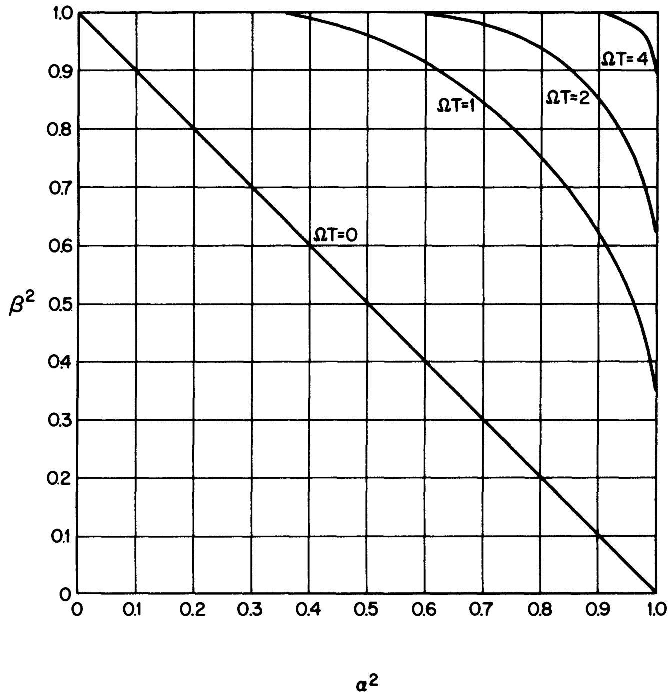

FIG. 3. *Possible combinations of* α^2 *and* β^2 *for different* ΩT. *Here* $\Omega = 2\pi W$. *Reprinted with permission from The Bell System Technical Journal. Copyright* 1961, *AT&T*.

both bandlimited and timelimited. The answer [2] is shown in Fig. 3 for several values of WT. The excluded region of the square $0 \leqq \alpha^2 \leqq 1$, $0 \leqq \beta^2 \leqq 1$ lies above the ellipse

$$\arccos \alpha + \arccos \beta = \arccos \sqrt{\lambda_0}.$$

ii) A bandlimited signal $r(t) \in B_W$ is observed on the interval $(-T/2, T/2)$. We wish to extend it beyond this interval. Since the $r_n(t)$ are complete in B_W,

$$r(t) = \sum_0^\infty c_n r_n(t).$$

Multiply by $r_m(t)$, integrate over the interval $(-T/2, T/2)$ and use (26) to find

$$c_n = \frac{1}{\lambda_n} \int_{-T/2}^{T/2} r_n(t) r(t)\, dt,$$

where only the observed values of $r(t)$ are used. The desired extension from $(-T/2, T/2)$ is then obtained as

$$r(t) = \sum_0^\infty r_n(t) \frac{1}{\lambda_n} \int_{-T/2}^{T/2} r(t') r_n(t')\, dt'.$$

See [1] for more on this problem.

On models. To describe one of the most important results arising from this work, I must digress again for a moment, this time to say a few words about models. Most of us in this audience are very aware of the distinction between the mathematical models we construct and study and the real-world phenomena that these mathematical structures are intended to explain or describe. Unfortunately, the engineer and the physicist sometimes overlook this distinction and attribute an objective reality to all the constructs of our models. This can lead to seemingly knotty paradoxes.

Our models of physical phenomena are merely games we play with symbols on paper, manipulating them according to well-defined rules. Certain quantities in our models will correspond, we hope, to observable measurable entities in the real-world situation we are attempting to describe. I call these *principal quantities* of the model. Almost always, however, there will be other quantities or constructs important to the model that have no counterpart in the real-world situation under study. I call these *secondary constructs* of the model. It is my contention that in useful, trustworthy models the principal quantities must be insensitive to small changes in the secondary constructs.

For example, in real laboratory notebooks the only numbers ever found are rational ones. The reading of meters or of dial settings in the real world always results in a finite string of symbols drawn from a finite list, usually a string of fewer than a dozen decimal digits. I do not see how a real measurement can yield an irrational number. Yet in our models both rationals and irrationals abound as we freely use the real number continuum. If a principal quantity of a model changes abruptly as we change some other quantity in the model from a rational value to an arbitrarily nearby irrational value, I would be very suspicious of the utility of the model. Rationality of quantities is a secondary construct of a model.

Continuity is also always a secondary construct. It makes no sense to ask whether the needle of a real voltmeter on a laboratory bench moves *continuously* with time in the sense that continuity is defined in mathematics. Observing the needle at millisecond intervals, or nanosecond intervals, comes no closer to answering that question than observing it at weekly intervals. Continuity is not a verifiable notion in the real world. However, it is useful in models. But in a satisfactory model, the introduction of tiny discontinuities into a continuous function should not change appreciably the principal quantities of the model—those parts of the model that we want to correspond to measurable real-world quantities.

And so it is with points at infinity and the detailed behavior of functions as their arguments become infinite. These are all generally secondary constructs of our models. They do not correspond to verifiable constructs in the real world. We introduce them for convenience in our mathematical game, but in modeling the real world we must see that principal quantities are properly insensitive to them.

The bandlimited signals and timelimited signals I have been discussing are cases in point. It is senseless to ask if real signals are bandlimited, or timelimited. Verification requires real measurements at arbitrarily high frequencies or at arbitrarily remote or future times, experiments that can never be carried out. The notions of bandlimitedness or timelimitedness belong to the engineer's model, not the real world. They are secondary constructs. As it suits him, he can assume in his model either that his signals are timelimited, or that they are bandlimited, or neither. But he should take care that the deductions he makes from his model about the real world do not depend sensitively on which assumption he has made.

The $2WT$-theorem. The "philosophy" I have just enunciated is discussed in more detail in my paper, *On Bandwidth* [8]. It ties in well with the spirit of the version of the $2WT$-theorem I now wish to discuss.

For years electrical engineers have espoused a folk theorem that asserts that if WT is large "the space of signals of duration T and bandwidth W has dimension $2WT$." Those who recognized that only the signal $r(t) \equiv 0$ has both finite duration T and finite bandwidth W preferred a version that stated: "for large WT, the space of signals of approximate duration T and approximate bandwidth W has approximate dimension $2WT$." With this amount of fuzzy hedging in its statement, the "theorem" is rightfully belittled by the mathematician. But there is truly substance behind these foggy notions; it has just been very difficult to clear away the mist and state this vague idea in a precise and useful way. To do so has been very important in setting limits to the rate at which one can communicate information in the real world [20].

Many different routes have led to this folk theorem. I mention two. i) A signal of duration T, say of support on $(0, T)$, has a Fourier series representation

$$r(t) = \sum_{0}^{\infty} a_n \cos 2\pi f_n t + \sum_{1}^{\infty} b_n \sin f_n t, \qquad f_n \equiv \frac{n}{T}, \tag{27}$$

valid for $0 < t < T$. Here f_n is the "frequency" of the elementary sinusoid $\cos 2\pi f_n t$ or $\sin 2\pi f_n t$ that is found in $r(t)$. If $r(t)$ "contains no frequencies higher than W," all terms in (27) must vanish when $f_n = n/T > W$, i.e. when $n > WT$. Then there are WT nonzero b's and $WT + 1$ nonzero a's in (27). Thus our "bandlimited and timelimited" $r(t)$ is specified by $2WT + 1 \sim 2WT$ constants.

ii) The Shannon-Whittaker sampling theorem [21] states that if $r(t) \in B_W$, then

$$r(t) = \sum_{-\infty}^{\infty} r(n/2W) \frac{\sin 2\pi W(t - n/2W)}{2\pi W(t - n/2W)}. \tag{28}$$

If now $r(t)$ vanishes outside the interval $(0, T)$, the nonzero terms in (28) occur only for $0 \leqq n/2W \leqq T$, i.e. for $n = 0, 1, 2, \cdots, [2WT]$. Thus the "space of bandlimited and timelimited functions" is asymptotically $2WT$-dimensional.

I will not demean you by pointing out what is fallacious in each of these arguments. But, if you will examine these "demonstrations" at your leisure, you will see that, in some sense, the lies I have told become less and less flagrant as either T becomes large in i) or W becomes large in ii). There is clearly some way in which this folk theorem is true.

To formulate it in a manner consonant with the philosophy of the preceding section, I will redefine the notions of bandlimited and timelimited so that they do not depend on the *detailed* behavior of signals or spectra at infinity. To this end, let us suppose that there is a smallest amount, say ε, of energy that we are able to detect by any means in the real world. In our model we now say that a *signal* $r(t)$ *is timelimited to the interval* $(-T/2, T/2)$ *at level* ε *if*

$$\int_{|t|>T/2} r(t)^2 \, dt < \varepsilon,$$

i.e., if the energy outside this time interval is less than we could measure in the laboratory. Similarly, in our model we way that a *signal is bandlimited with bandwidth* W *at level* ε if

$$\int_{|f|>W} |R(f)|^2 \, df < \varepsilon,$$

i.e., if the energy lying outside the frequency range is less than we can measure.

These notions of signals bandlimited and timelimited at a level ε are really quite different from the notions I spoke of earlier and which I henceforth refer to as *absolutely timelimited* or *absolutely bandlimited* signals. Now every signal is both timelimited *and* bandlimited at level ε (for some T and some W), whereas we saw before that only the always-zero signal is absolutely timelimited and absolutely bandlimited. The smallest value of W for which $r(t)$ is bandlimited at level ε is called its *bandwidth at level* ε. Note that changing a signal by a multiplicative scale factor in general changes its bandwidth at level ε. The bandwidth of a signal absolutely bandlimited is uneffected by such scale changes.

A final definition is needed in order to state the $2WT$-theorem. *A set F of signals has approximate dimension N at level* ε *during the interval* $(-T/2, T/2)$ *if there is a set of* $N = N(T, \varepsilon)$ *signals* $\phi_1, \phi_2, \cdots, \phi_N$ *such that for each* $r(t) \in F$ *there exists* $a_1, a_2, \cdots, a_N$ *such that*

$$\int_{-T/2}^{T/2} \left[r(t) - \sum_1^N a_j \phi_j(t) \right]^2 dt < \varepsilon$$

and there is no set of $N - 1$ *functions that will approximate every* $r \in F$ *in this manner.* Stated otherwise, every signal in F can be so well approximated during $-T/2 < t < T/2$ by a signal in the linear span of $\phi_1, \phi_2, \cdots, \phi_N$ that in the real world we could not detect the difference between the signal and its approximation.

The $2WT$-theorem now reads:

Let F_ε *be the set of signals timelimited to* $(-T/2, T/2)$ *at level* ε *and bandlimited to* $(-W, W)$ *at level* ε. *Let* $N(W, T, \varepsilon, \varepsilon')$ *be the approximate dimension of* F_ε *at level* ε′. *Then for every* $\varepsilon' > \varepsilon$,

$$\lim_{T\to\infty} \frac{N(W, T, \varepsilon, \varepsilon')}{T} = 2W, \qquad \lim_{W\to\infty} \frac{N(W, T, \varepsilon, \varepsilon')}{W} = 2T. \tag{29}$$

Note that the result (29) does not really depend on ε. If at some later date we can measure much smaller quantities of energy than we can today, still asymptotically the set of signals which in the real world we must consider to be limited to duration T and bandwidth W will be only $2WT$-dimensional.

Proof of this theorem can be found in [8]. It depends heavily on properties of the pswf's $\psi_0, \psi_1, \cdots$ and the eigenvalues $\lambda_0, \lambda_1, \cdots$.

Concluding remarks. In closing I want to point out a few of the generalizations of these ideas that exhibit much of the same structure. First of all, there is the obvious extension of the concentration problem to functions of several variables [4]. The two-dimensional signal $r(t_1, t_2)$ *has bandwidth* W if its Fourier transform

$$R(f_1, f_2) \equiv \int_{-\infty}^{\infty} dt_1 \int_{-\infty}^{\infty} dt_2 e^{2\pi i(t_1 f_1 + t_2 f_2)} r(t_1, t_2)$$

vanishes for $f_1^2 + f_2^2 > W^2$. How large can the quantity

$$\alpha^2 = \iint_{t_1^2 + t_2^2 \leq T^2} r^2(t_1, t_2)\, dt_1 dt_2 \Big/ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} r^2(t_1, t_2)\, dt_1 dt_2$$

be for signals of bandwidth W? The problem eventually works out much like the 1-dimensional case. First introduce polar coordinates $f_1 = \rho \cos \phi$, $f_2 = \rho \sin \phi$. Then write

$$R(f_1, f_2) = \frac{1}{\sqrt{\rho}} \sum_N \psi_N(\rho) e^{iN\phi}.$$

One finds (see [4] for details) that ψ_N must satisfy

$$\int_0^1 J_N(cxy) \sqrt{cxy}\, \psi_N(y)\, dy = \lambda \psi_N(x), \qquad 0 \leqq x \leqq 1$$

where $J_N(x)$ is the usual Bessel function and now $c = 2\pi WT$. Further progress can be made since we have the lucky accident that this eigenvalue problem has the same eigenfunctions as the differential equation problem

$$\frac{d}{dx}(1 - x^2)\frac{d\psi}{dx} + \left(\chi - c^2x^2 + \frac{1/4 - N^2}{x^2}\right)\psi = 0.$$

Double orthogonality and many other properties carry over. The work generalizes easily to functions of $D > 2$ variables. By luck, the case $D = 2$ is of importance in the theory of lasers [22].

Generalizing in another way, the *amplitude spectrum* of a double infinite L^2 sequence $\{r_n\} = \cdots, r_{-1}, r_0, r_1, \cdots$ is defined as

$$R(f) \equiv \sum_{-\infty}^{\infty} r_n e^{2\pi i n f}, \qquad |f| \leqq \frac{1}{2}.$$

The sequence $\{r_n\}$ is *bandlimited with bandwidth* $W < 1/2$ if $R(f)$ vanishes for $1/2 \geqq |f| > W$. The concentration problem for these bandlimited sequences looks much the same [10]. There is an integral equation formulation and a differential equation formulation, and in addition, here there is also a difference equation approach. The solutions are called *discrete prolate spheroidal sequences;* their amplitude spectra are *discrete prolate spheroidal functions.*

Still another discrete version comes from considering the concentration problem for a finite sequence and its discrete Fourier transform [12]. The general structure is unchanged.

E. N. Gilbert and I [9] considered a related problem. We sought the nth degree polynomial, $f(x)$, which maximizes

$$\alpha^2 \equiv \int_{I_a} |f(x)|^2\,dx \Big/ \int_{I_b} |f(x)|^2\,dx$$

where I_a and I_b are intervals on the real x-axis. A family of doubly orthogonal concentrated polynomials results. There is an integral equation formulation of the problem and, in certain cases, a differential equation formulation. These cases are when I_a and I_b are adjacent, and when I_a is centrally located within I_b. With the intervals related otherwise, we could not find a differential equation formulation. An analogous dependence of the differential equation formulation of these problems on the nature of the regions of concentration had been noted very early [23].

Other generalizations are to be found in [11], [13]–[15].

In closing I wish to call attention to two valuable papers by my friend and colleague H. J. Landau [24], [25]. The first of these examines part of the theory in a very general setting. The second is a brilliant application of the theory to a problem that has received much attention in the engineering literature—sampling and the Nyquist rate.

Many other references to applications of the work can be found in the papers cited here.

REFERENCES

[1] D. Slepian and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty,* I, Bell Syst. Tech. J., 40 (1961), pp. 43–64.

[2] H. J. Landau and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty,* II, Bell Syst. Tech. J., 40 (1961), pp. 65–84.

[3] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*, III, Bell Syst. Tech. J., 41 (1962), pp. 1295–1336.

[4] DAVID SLEPIAN, *Prolate spheroidal wave functions, Fourier analysis and uncertainty—*IV: *Extensions to many dimensions; Generalized prolate spheroidal functions*, Bell Syst. Tech. J., 43 (1964), pp. 3009–3058.

[5] DAVID SLEPIAN AND ESTELLE SONNENBLICK, *Eigenvalues associated with prolate spheroidal wave functions of zero order*, Bell Syst. Tech. J., 44 (1965), pp. 1745–1760.

[6] DAVID SLEPIAN, *Some asymptotic expansions for prolate spheroidal wave functions*, J. Math. and Physics, 44 (1965), pp. 99–140.

[7] ———, *Analytic solution of two apodization problems*, J. Opt. Soc. Am., 55 (1965), pp. 1110–1115.

[8] ———, *On bandwidth*, Proc. IEEE, 64 (1976), pp. 292–300.

[9] E. N. GILBERT AND D. SLEPIAN, *Doubly orthogonal concentrated polynomials*, SIAM J. Math. Anal. 8 (1977), pp. 290–319.

[10] D. SLEPIAN, *Prolate spheroidal wave functions, Fourier analysis and uncertainty—*V: *The discrete case*, Bell Syst. Tech. J., 57 (1978), pp. 1371–1430.

[11] F. A. GRUNBAUM, *Toeplitz matrices commuting with tridiagonal matrices*, J. Linear Alg. and Appl., 40 (1981), pp. 25–36.

[12] ———, *Eigenvectors of a Toeplitz matrix: Discrete version of the prolate spheroidal wave functions*, SIAM J. Alg. Disc. Meth., 2 (1981), pp. 136–141.

[13] F. A. GRUNBAUM, L. LONGHI AND M. PERLSTADT, *Differential operators commuting with finite convolution integral operators: some non-Abelian examples*, SIAM J. Appl. Math., 42 (1982), pp. 941–985.

[14] M. PERLSTADT, *Chopped orthogonal polynomial expansions—some discrete cases*, SIAM J. Alg. Disc. Meth., 4 (1983), pp. 94–100.

[15] F. A. GRUNBAUM, *A new property of a reproducing kernel for classical orthogonal polynomials*, J. Math. Anal. Appl., to appear.

[16] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. 1, Interscience Publishers, New York, 1955, pp. 122–134.

[17] J. MEIXNER AND F. W. SCHÄFKE, *Mathieusche Funktionen und Spharoid-functionen*, Julius Springer, Berlin, 1954.

[18] C. FLAMMER, *Spheroidal Wave Functions*, Stanford Univ. Press, Stanford, CA, 1957.

[19] H. J. LANDAU AND H. WIDOM, *The eigenvalue distribution of time and frequency limiting*, J. Math. Anal. Appl., 77 (1980), pp. 469–481.

[20] A. D. WYNER, *Capacity of the band-limited Gaussian channel*, Bell Syst. Tech. J., 45 (1965), pp. 359–395; *A note on the capacity of the band-limited Gaussian channel*, Bell Syst. Tech. J., 55 (1976), pp. 343–346.

[21] H. DYM AND H. P. MCKEAN, *Fourier Series and Integrals*, Academic Press, New York, 1972, p. 129.

[22] H. KOGELNIK AND T. LI, *Laser beams and resonators*, Proc. IEEE, 54 (1966), pp. 1312–1329.

[23] J. A. MORRISON, *On the eigenfunctions corresponding to the bandpass kernel in the case of degeneracy*, Quart. Appl. Math., 21 (1963), pp. 13–19.

[24] H. J. LANDAU, *Necessary density conditions for sampling and interpolation of certain entire functions*, Acta Math., 117 (1967), pp. 37–52.

[25] ———, *Sampling, data transmission, and the Nyquist rate*, Proc. IEEE, 55 (1967), pp. 1701–1706.