Chapter 12.   Fast Fourier Transform

12.0 Introduction

A very large class of important computational problems falls under the general rubric of "Fourier transform methods" or "spectral methods." For some of these problems, the Fourier transform is simply an efficient computational tool for accomplishing certain common manipulations of data. In other cases, we have problems for which the Fourier transform (or the related "power spectrum") is itself of intrinsic interest. These two kinds of problems share a common methodology.

Largely for historical reasons the literature on Fourier and spectral methods has been disjoint from the literature on "classical" numerical analysis. Nowadays there is no justification for such a split. Fourier methods are commonplace in research and we shall not treat them as specialized or arcane. At the same time, we realize that many computer users have had relatively less experience with this field than with, say, differential equations or numerical integration. Therefore our summary of analytical results will be more complete. Numerical algorithms, per se, begin in §12.2. Various applications of Fourier transform methods are discussed in Chapter 13.

A physical process can be described either in the *time domain*, by the values of some quantity h as a function of time t, e.g., $h(t)$, or else in the *frequency domain*, where the process is specified by giving its amplitude H (generally a complex number indicating phase also) as a function of frequency f, that is $H(f)$, with $-\infty < f < \infty$. For many purposes it is useful to think of $h(t)$ and $H(f)$ as being two different *representations* of the *same* function. One goes back and forth between these two representations by means of the *Fourier transform* equations,

$$
\begin{aligned}
H(f) &= \int_{-\infty}^{\infty} h(t)e^{2\pi i f t} dt \\
h(t) &= \int_{-\infty}^{\infty} H(f)e^{-2\pi i f t} df
\end{aligned}
\tag{12.0.1}
$$

If t is measured in seconds, then f in equation (12.0.1) is in cycles per second, or Hertz (the unit of frequency). However, the equations work with other units too. If h is a function of position x (in meters), H will be a function of inverse wavelength (cycles per meter), and so on. If you are trained as a physicist or mathematician, you are probably more used to using *angular frequency* ω, which is given in *radians* per sec. The relation between ω and f, $H(\omega)$ and $H(f)$ is

$$
\omega \equiv 2\pi f \qquad H(\omega) \equiv [H(f)]_{f=\omega/2\pi}
\tag{12.0.2}
$$

and equation (12.0.1) looks like this

$$H(\omega) = \int_{-\infty}^{\infty} h(t)e^{i\omega t}dt$$

$$h(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(\omega)e^{-i\omega t}d\omega \qquad (12.0.3)$$

We were raised on the ω-convention, but we changed! There are fewer factors of 2π to remember if you use the f-convention, especially when we get to discretely sampled data in §12.1.

From equation (12.0.1) it is evident at once that Fourier transformation is a *linear* operation. The transform of the sum of two functions is equal to the sum of the transforms. The transform of a constant times a function is that same constant times the transform of the function.

In the time domain, function $h(t)$ may happen to have one or more special symmetries It might be *purely real* or *purely imaginary* or it might be *even*, $h(t) = h(-t)$, or *odd*, $h(t) = -h(-t)$. In the frequency domain, these symmetries lead to relationships between $H(f)$ and $H(-f)$. The following table gives the correspondence between symmetries in the two domains:

If ...	then ...
$h(t)$ is real	$H(-f) = [H(f)]^*$
$h(t)$ is imaginary	$H(-f) = -[H(f)]^*$
$h(t)$ is even	$H(-f) = H(f)$ [i.e., $H(f)$ is even]
$h(t)$ is odd	$H(-f) = -H(f)$ [i.e., $H(f)$ is odd]
$h(t)$ is real and even	$H(f)$ is real and even
$h(t)$ is real and odd	$H(f)$ is imaginary and odd
$h(t)$ is imaginary and even	$H(f)$ is imaginary and even
$h(t)$ is imaginary and odd	$H(f)$ is real and odd

In subsequent sections we shall see how to use these symmetries to increase computational efficiency.

Here are some other elementary properties of the Fourier transform. (We'll use the "$\Longleftrightarrow$" symbol to indicate transform pairs.) If

$$h(t) \Longleftrightarrow H(f)$$

is such a pair, then other transform pairs are

$$h(at) \Longleftrightarrow \frac{1}{|a|}H(\frac{f}{a}) \qquad \text{"time scaling"} \qquad (12.0.4)$$

$$\frac{1}{|b|}h(\frac{t}{b}) \Longleftrightarrow H(bf) \qquad \text{"frequency scaling"} \qquad (12.0.5)$$

$$h(t - t_0) \Longleftrightarrow H(f)\, e^{2\pi i f t_0} \qquad \text{"time shifting"} \qquad (12.0.6)$$

$$h(t)\, e^{-2\pi i f_0 t} \Longleftrightarrow H(f - f_0) \qquad \text{"frequency shifting"} \qquad (12.0.7)$$

With two functions $h(t)$ and $g(t)$, and their corresponding Fourier transforms $H(f)$ and $G(f)$, we can form two combinations of special interest. The *convolution* of the two functions, denoted $g * h$, is defined by

$$g * h \equiv \int_{-\infty}^{\infty} g(\tau)h(t - \tau) \, d\tau \qquad (12.0.8)$$

Note that $g * h$ is a function in the time domain and that $g * h = h * g$. It turns out that the function $g * h$ is one member of a simple transform pair

$$g * h \Longleftrightarrow G(f)H(f) \qquad \text{"Convolution Theorem"} \qquad (12.0.9)$$

In other words, the Fourier transform of the convolution is just the product of the individual Fourier transforms.

The *correlation* of two functions, denoted $\text{Corr}(g, h)$, is defined by

$$\text{Corr}(g, h) \equiv \int_{-\infty}^{\infty} g(\tau + t)h(\tau) \, d\tau \qquad (12.0.10)$$

The correlation is a function of t, which is called the *lag*. It therefore lies in the time domain, and it turns out to be one member of the transform pair:

$$\text{Corr}(g, h) \Longleftrightarrow G(f)H^*(f) \qquad \text{"Correlation Theorem"} \qquad (12.0.11)$$

[More generally, the second member of the pair is $G(f)H(-f)$, but we are restricting ourselves to the usual case in which g and h are real functions, so we take the liberty of setting $H(-f) = H^*(f)$.] This result shows that multiplying the Fourier transform of one function by the complex conjugate of the Fourier transform of the other gives the Fourier transform of their correlation. The correlation of a function with itself is called its *autocorrelation*. In this case (12.0.11) becomes the transform pair

$$\text{Corr}(g, g) \Longleftrightarrow |G(f)|^2 \qquad \text{"Wiener-Khinchin Theorem"} \qquad (12.0.12)$$

The *total power* in a signal is the same whether we compute it in the time domain or in the frequency domain. This result is known as *Parseval's theorem*:

$$\text{Total Power} \equiv \int_{-\infty}^{\infty} |h(t)|^2 \, dt = \int_{-\infty}^{\infty} |H(f)|^2 \, df \qquad (12.0.13)$$

Frequently one wants to know "how much power" is contained in the frequency interval between f and $f + df$. In such circumstances one does not usually distinguish between positive and negative f, but rather regards f as varying from 0 ("zero frequency" or D.C.) to $+\infty$. In such cases, one defines the *one-sided power spectral density (PSD)* of the function h as

$$P_h(f) \equiv |H(f)|^2 + |H(-f)|^2 \qquad 0 \leq f < \infty \qquad (12.0.14)$$

so that the total power is just the integral of $P_h(f)$ from $f = 0$ to $f = \infty$. When the function $h(t)$ is real, then the two terms in (12.0.14) are equal, so $P_h(f) = 2\,|H(f)|^2$.
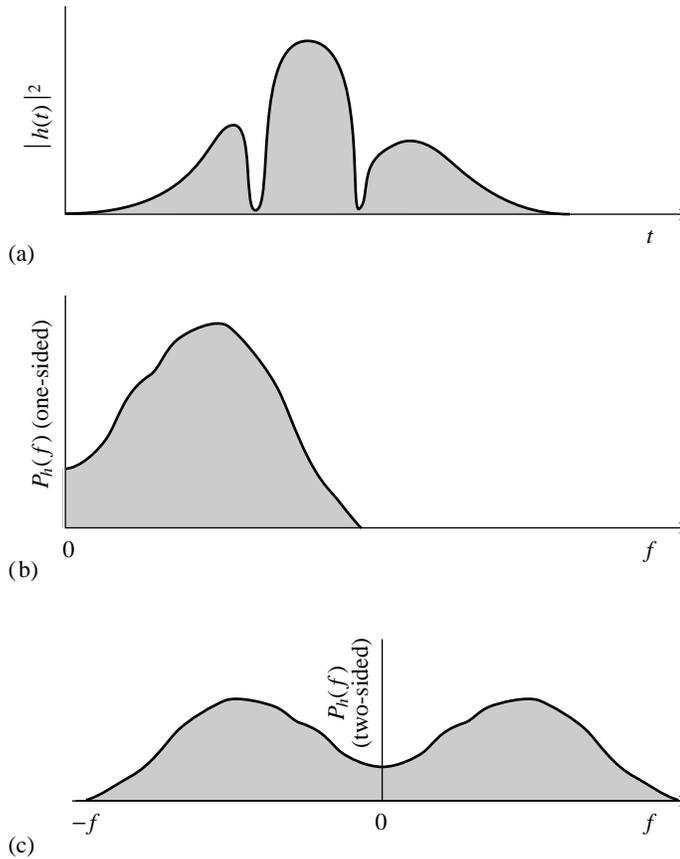
Figure 12.0.1. Normalizations of one- and two-sided power spectra. The area under the square of the function, (a), equals the area under its one-sided power spectrum at positive frequencies, (b), and also equals the area under its two-sided power spectrum at positive and negative frequencies, (c).

Be warned that one occasionally sees PSDs defined without this factor two. These, strictly speaking, are called *two-sided power spectral densities*, but some books are not careful about stating whether one- or two-sided is to be assumed. We will always use the one-sided density given by equation (12.0.14). Figure 12.0.1 contrasts the two conventions.

If the function $h(t)$ goes endlessly from $-\infty < t < \infty$, then its total power and power spectral density will, in general, be infinite. Of interest then is the *(one- or two-sided) power spectral density per unit time*. This is computed by taking a long, but finite, stretch of the function $h(t)$, computing its PSD [that is, the PSD of a function that equals $h(t)$ in the finite stretch but is zero everywhere else], and then dividing the resulting PSD by the length of the stretch used. Parseval's theorem in this case states that the integral of the one-sided PSD-per-unit-time over positive frequency is equal to the mean square amplitude of the signal $h(t)$.

You might well worry about how the PSD-per-unit-time, which is a function of frequency f, converges as one evaluates it using longer and longer stretches of data. This interesting question is the content of the subject of "power spectrum estimation," and will be considered below in §13.4–§13.7. A crude answer for

now is: The PSD-per-unit-time converges to finite values at all frequencies *except* those where $h(t)$ has a discrete sine-wave (or cosine-wave) component of finite amplitude. At those frequencies, it becomes a delta-function, i.e., a sharp spike, whose width gets narrower and narrower, but whose area converges to be the mean square amplitude of the discrete sine or cosine component at that frequency.

We have by now stated all of the analytical formalism that we will need in this chapter with one exception: In computational work, especially with experimental data, we are almost never given a continuous function $h(t)$ to work with, but are given, rather, a list of measurements of $h(t_i)$ for a discrete set of t_i's. The profound implications of this seemingly unimportant fact are the subject of the next section.

CITED REFERENCES AND FURTHER READING:

Champeney, D.C. 1973, *Fourier Transforms and Their Physical Applications* (New York: Academic Press).

Elliott, D.F., and Rao, K.R. 1982, *Fast Transforms: Algorithms, Analyses, Applications* (New York: Academic Press).

12.1 Fourier Transform of Discretely Sampled Data

In the most common situations, function $h(t)$ is sampled (i.e., its value is recorded) at evenly spaced intervals in time. Let Δ denote the time interval between consecutive samples, so that the sequence of sampled values is

$$h_n = h(n\Delta) \qquad n = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots \qquad (12.1.1)$$

The reciprocal of the time interval Δ is called the *sampling rate*; if Δ is measured in seconds, for example, then the sampling rate is the number of samples recorded per second.

Sampling Theorem and Aliasing

For any sampling interval Δ, there is also a special frequency f_c, called the *Nyquist critical frequency*, given by

$$f_c \equiv \frac{1}{2\Delta} \qquad (12.1.2)$$

If a sine wave of the Nyquist critical frequency is sampled at its positive peak value, then the next sample will be at its negative trough value, the sample after that at the positive peak again, and so on. Expressed otherwise: *Critical sampling of a sine wave is two sample points per cycle.* One frequently chooses to measure time in units of the sampling interval Δ. In this case the Nyquist critical frequency is just the constant 1/2.

The Nyquist critical frequency is important for two related, but distinct, reasons. One is good news, and the other bad news. First the good news. It is the remarkable

fact known as the *sampling theorem*: If a continuous function $h(t)$, sampled at an interval Δ, happens to be *bandwidth limited* to frequencies smaller in magnitude than f_c, i.e., if $H(f) = 0$ for all $|f| \geq f_c$, then the function $h(t)$ is *completely determined* by its samples h_n. In fact, $h(t)$ is given explicitly by the formula

$$h(t) = \Delta \sum_{n=-\infty}^{+\infty} h_n \frac{\sin[2\pi f_c(t - n\Delta)]}{\pi(t - n\Delta)} \qquad (12.1.3)$$

This is a remarkable theorem for many reasons, among them that it shows that the "information content" of a bandwidth limited function is, in some sense, infinitely smaller than that of a general continuous function. Fairly often, one is dealing with a signal that is known on physical grounds to be bandwidth limited (or at least approximately bandwidth limited). For example, the signal may have passed through an amplifier with a known, finite frequency response. In this case, the sampling theorem tells us that the entire information content of the signal can be recorded by sampling it at a rate Δ^{-1} equal to twice the maximum frequency passed by the amplifier (cf. 12.1.2).

Now the bad news. The bad news concerns the effect of sampling a continuous function that is *not* bandwidth limited to less than the Nyquist critical frequency. In that case, it turns out that all of the power spectral density that lies outside of the frequency range $-f_c < f < f_c$ is spuriously moved into that range. This phenomenon is called *aliasing*. Any frequency component outside of the frequency range $(-f_c, f_c)$ is *aliased* (falsely translated) into that range by the very act of discrete sampling. You can readily convince yourself that two waves $\exp(2\pi i f_1 t)$ and $\exp(2\pi i f_2 t)$ give the same samples at an interval Δ if and only if f_1 and f_2 differ by a multiple of $1/\Delta$, which is just the width in frequency of the range $(-f_c, f_c)$. There is little that you can do to remove aliased power once you have discretely sampled a signal. The way to overcome aliasing is to (i) know the natural bandwidth limit of the signal — or else enforce a known limit by analog filtering of the continuous signal, and then (ii) sample at a rate sufficiently rapid to give at least two points per cycle of the highest frequency present. Figure 12.1.1 illustrates these considerations.

To put the best face on this, we can take the alternative point of view: If a continuous function has been competently sampled, then, when we come to estimate its Fourier transform from the discrete samples, we can *assume* (or rather we *might as well* assume) that its Fourier transform is equal to zero outside of the frequency range in between $-f_c$ and f_c. Then we look to the Fourier transform to tell whether the continuous function *has* been competently sampled (aliasing effects minimized). We do this by looking to see whether the Fourier transform is already approaching zero as the frequency approaches f_c from below, or $-f_c$ from above. If, on the contrary, the transform is going towards some finite value, then chances are that components outside of the range have been folded back over onto the critical range.

Discrete Fourier Transform

We now estimate the Fourier transform of a function from a finite number of its sampled points. Suppose that we have N consecutive sampled values

$$h_k \equiv h(t_k), \qquad t_k \equiv k\Delta, \qquad k = 0, 1, 2, \ldots, N - 1 \qquad (12.1.4)$$

(a)

(b)

(c)
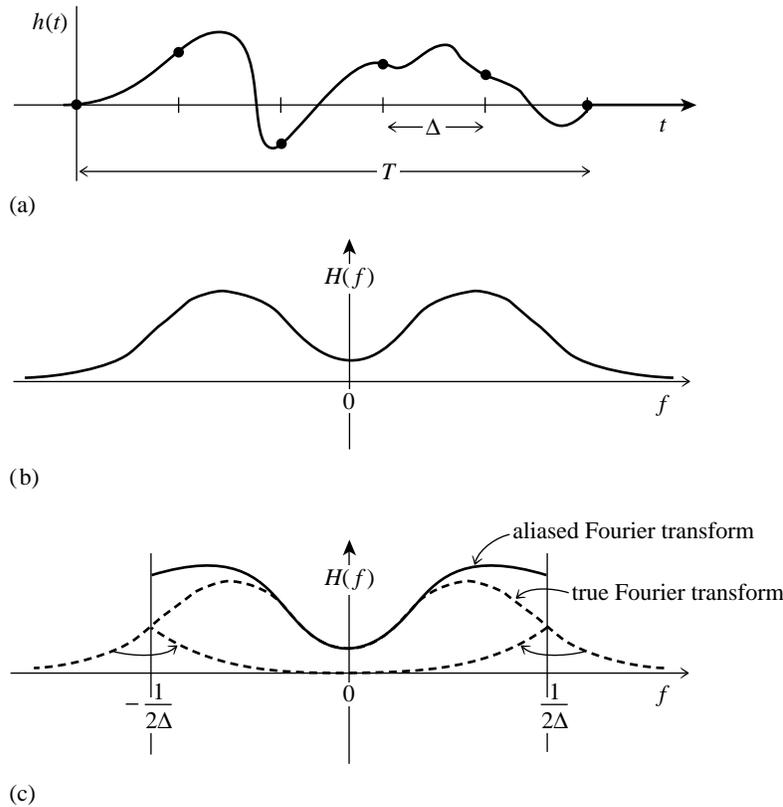
Figure 12.1.1.   The continuous function shown in (a) is nonzero only for a finite interval of time T. It follows that its Fourier transform, whose modulus is shown schematically in (b), is not bandwidth limited but has finite amplitude for all frequencies. If the original function is sampled with a sampling interval Δ, as in (a), then the Fourier transform (c) is defined only between plus and minus the Nyquist critical frequency. Power outside that range is folded over or "aliased" into the range. The effect can be eliminated only by low-pass filtering the original function *before sampling*.

so that the sampling interval is Δ. To make things simpler, let us also suppose that N is even. If the function $h(t)$ is nonzero only in a finite interval of time, then that whole interval of time is supposed to be contained in the range of the N points given. Alternatively, if the function $h(t)$ goes on forever, then the sampled points are supposed to be at least "typical" of what $h(t)$ looks like at all other times.

With N numbers of input, we will evidently be able to produce no more than N independent numbers of output. So, instead of trying to estimate the Fourier transform $H(f)$ at all values of f in the range $-f_c$ to f_c, let us seek estimates only at the discrete values

$$f_n \equiv \frac{n}{N\Delta}, \qquad n = -\frac{N}{2}, \ldots, \frac{N}{2} \qquad (12.1.5)$$

The extreme values of n in (12.1.5) correspond exactly to the lower and upper limits of the Nyquist critical frequency range. If you are really on the ball, you will have noticed that there are $N + 1$, not N, values of n in (12.1.5); it will turn out that the two extreme values of n are not independent (in fact they are equal), but all the others are. This reduces the count to N.

The remaining step is to approximate the integral in (12.0.1) by a discrete sum:

$$H(f_n) = \int_{-\infty}^{\infty} h(t) e^{2\pi i f_n t} dt \approx \sum_{k=0}^{N-1} h_k\, e^{2\pi i f_n t_k} \Delta = \Delta \sum_{k=0}^{N-1} h_k\, e^{2\pi i k n/N} \tag{12.1.6}$$

Here equations (12.1.4) and (12.1.5) have been used in the final equality. The final summation in equation (12.1.6) is called the *discrete Fourier transform* of the N points h_k. Let us denote it by H_n,

$$H_n \equiv \sum_{k=0}^{N-1} h_k\, e^{2\pi i k n/N} \tag{12.1.7}$$

The discrete Fourier transform maps N complex numbers (the h_k's) into N complex numbers (the H_n's). It does not depend on any dimensional parameter, such as the time scale Δ. The relation (12.1.6) between the discrete Fourier transform of a set of numbers and their continuous Fourier transform when they are viewed as samples of a continuous function sampled at an interval Δ can be rewritten as

$$H(f_n) \approx \Delta H_n \tag{12.1.8}$$

where f_n is given by (12.1.5).

Up to now we have taken the view that the index n in (12.1.7) varies from $-N/2$ to $N/2$ (cf. 12.1.5). You can easily see, however, that (12.1.7) is periodic in n, with period N. Therefore, $H_{-n} = H_{N-n}$ $n = 1, 2, \ldots$. With this conversion in mind, one generally lets the n in H_n vary from 0 to $N - 1$ (one complete period). Then n and k (in h_k) vary exactly over the same range, so the mapping of N numbers into N numbers is manifest. When this convention is followed, you must remember that zero frequency corresponds to $n = 0$, positive frequencies $0 < f < f_c$ correspond to values $1 \le n \le N/2 - 1$, while negative frequencies $-f_c < f < 0$ correspond to $N/2 + 1 \le n \le N - 1$. The value $n = N/2$ corresponds to *both* $f = f_c$ and $f = -f_c$.

The discrete Fourier transform has symmetry properties almost exactly the same as the continuous Fourier transform. For example, all the symmetries in the table following equation (12.0.3) hold if we read h_k for $h(t)$, H_n for $H(f)$, and H_{N-n} for $H(-f)$. (Likewise, "even" and "odd" in time refer to whether the values h_k at k and $N - k$ are identical or the negative of each other.)

The formula for the discrete *inverse* Fourier transform, which recovers the set of h_k's exactly from the H_n's is:

$$h_k = \frac{1}{N} \sum_{n=0}^{N-1} H_n\, e^{-2\pi i k n/N} \tag{12.1.9}$$

Notice that the only differences between (12.1.9) and (12.1.7) are (i) changing the sign in the exponential, and (ii) dividing the answer by N. This means that a routine for calculating discrete Fourier transforms can also, with slight modification, calculate the inverse transforms.

The discrete form of Parseval's theorem is

$$\sum_{k=0}^{N-1} |h_k|^2 = \frac{1}{N} \sum_{n=0}^{N-1} |H_n|^2 \qquad (12.1.10)$$

There are also discrete analogs to the convolution and correlation theorems (equations 12.0.9 and 12.0.11), but we shall defer them to §13.1 and §13.2, respectively.

CITED REFERENCES AND FURTHER READING:

Brigham, E.O. 1974, *The Fast Fourier Transform* (Englewood Cliffs, NJ: Prentice-Hall).

Elliott, D.F., and Rao, K.R. 1982, *Fast Transforms: Algorithms, Analyses, Applications* (New York: Academic Press).

12.2 Fast Fourier Transform (FFT)

How much computation is involved in computing the discrete Fourier transform (12.1.7) of N points? For many years, until the mid-1960s, the standard answer was this: Define W as the complex number

$$W \equiv e^{2\pi i/N} \qquad (12.2.1)$$

Then (12.1.7) can be written as

$$H_n = \sum_{k=0}^{N-1} W^{nk} h_k \qquad (12.2.2)$$

In other words, the vector of h_k's is multiplied by a matrix whose (n, k)th element is the constant W to the power $n \times k$. The matrix multiplication produces a vector result whose components are the H_n's. This matrix multiplication evidently requires N^2 complex multiplications, plus a smaller number of operations to generate the required powers of W. So, the discrete Fourier transform appears to be an $O(N^2)$ process. These appearances are deceiving! The discrete Fourier transform can, in fact, be computed in $O(N \log_2 N)$ operations with an algorithm called the *fast Fourier transform*, or *FFT*. The difference between $N \log_2 N$ and N^2 is immense. With $N = 10^6$, for example, it is the difference between, roughly, 30 seconds of CPU time and 2 weeks of CPU time on a microsecond cycle time computer. The existence of an FFT algorithm became generally known only in the mid-1960s, from the work of J.W. Cooley and J.W. Tukey. Retrospectively, we now know (see [1]) that efficient methods for computing the DFT had been independently discovered, and in some cases implemented, by as many as a dozen individuals, starting with Gauss in 1805!

One "rediscovery" of the FFT, that of Danielson and Lanczos in 1942, provides one of the clearest derivations of the algorithm. Danielson and Lanczos showed that a discrete Fourier transform of length N can be rewritten as the sum of two discrete Fourier transforms, each of length $N/2$. One of the two is formed from the

even-numbered points of the original N, the other from the odd-numbered points. The proof is simply this:

$$
\begin{aligned}
F_k &= \sum_{j=0}^{N-1} e^{2\pi ijk/N} f_j \\
&= \sum_{j=0}^{N/2-1} e^{2\pi ik(2j)/N} f_{2j} + \sum_{j=0}^{N/2-1} e^{2\pi ik(2j+1)/N} f_{2j+1} \\
&= \sum_{j=0}^{N/2-1} e^{2\pi ikj/(N/2)} f_{2j} + W^k \sum_{j=0}^{N/2-1} e^{2\pi ikj/(N/2)} f_{2j+1} \\
&= F_k^e + W^k\, F_k^o
\end{aligned}
\tag{12.2.3}
$$

In the last line, W is the same complex constant as in (12.2.1), F_k^e denotes the kth component of the Fourier transform of length $N/2$ formed from the even components of the original f_j's, while F_k^o is the corresponding transform of length $N/2$ formed from the odd components. Notice also that k in the last line of (12.2.3) varies from 0 to N, not just to $N/2$. Nevertheless, the transforms F_k^e and F_k^o are periodic in k with length $N/2$. So each is repeated through two cycles to obtain F_k.

The wonderful thing about the *Danielson-Lanczos Lemma* is that it can be used recursively. Having reduced the problem of computing F_k to that of computing F_k^e and F_k^o, we can do the same reduction of F_k^e to the problem of computing the transform of *its* $N/4$ even-numbered input data and $N/4$ odd-numbered data. In other words, we can define F_k^{ee} and F_k^{eo} to be the discrete Fourier transforms of the points which are respectively even-even and even-odd on the successive subdivisions of the data.

Although there are ways of treating other cases, by far the easiest case is the one in which the original N is an integer power of 2. In fact, we categorically recommend that you *only* use FFTs with N a power of two. If the length of your data set is not a power of two, pad it with zeros up to the next power of two. (We will give more sophisticated suggestions in subsequent sections below.) With this restriction on N, it is evident that we can continue applying the Danielson-Lanczos Lemma until we have subdivided the data all the way down to transforms of length 1. What is the Fourier transform of length one? It is just the identity operation that copies its one input number into its one output slot! In other words, for every pattern of $\log_2 N$ e's and o's, there is a one-point transform that is just one of the input numbers f_n

$$
F_k^{eoeeoeo\cdots oee} = f_n \qquad \text{for some } n
\tag{12.2.4}
$$

(Of course this one-point transform actually does not depend on k, since it is periodic in k with period 1.)

The next trick is to figure out which value of n corresponds to which pattern of e's and o's in equation (12.2.4). The answer is: Reverse the pattern of e's and o's, then let $e = 0$ and $o = 1$, and you will have, *in binary* the value of n. Do you see why it works? It is because the successive subdivisions of the data into even and odd are tests of successive low-order (least significant) bits of n. This idea of *bit reversal* can be exploited in a very clever way which, along with the Danielson-Lanczos

Figure 12.2.1.   Reordering an array (here of length 8) by bit reversal, (a) between two arrays, versus (b) in place. Bit reversal reordering is a necessary part of the fast Fourier transform (FFT) algorithm.

Lemma, makes FFTs practical: Suppose we take the original vector of data f_j and rearrange it into bit-reversed order (see Figure 12.2.1), so that the individual numbers are in the order not of j, but of the number obtained by bit-reversing j. Then the bookkeeping on the recursive application of the Danielson-Lanczos Lemma becomes extraordinarily simple. The points as given are the one-point transforms. We combine adjacent pairs to get two-point transforms, then combine adjacent pairs of pairs to get 4-point transforms, and so on, until the first and second halves of the whole data set are combined into the final transform. Each combination takes of order N operations, and there are evidently $\log_2 N$ combinations, so the whole algorithm is of order $N \log_2 N$ (assuming, as is the case, that the process of sorting into bit-reversed order is no greater in order than $N \log_2 N$).

This, then, is the structure of an FFT algorithm: It has two sections. The first section sorts the data into bit-reversed order. Luckily this takes no additional storage, since it involves only swapping pairs of elements. (If k_1 is the bit reverse of k_2, then k_2 is the bit reverse of k_1.) The second section has an outer loop that is executed $\log_2 N$ times and calculates, in turn, transforms of length $2, 4, 8, \ldots, N$. For each stage of this process, two nested inner loops range over the subtransforms already computed and the elements of each transform, implementing the Danielson-Lanczos Lemma. The operation is made more efficient by restricting external calls for trigonometric sines and cosines to the outer loop, where they are made only $\log_2 N$ times. Computation of the sines and cosines of multiple angles is through simple recurrence relations in the inner loops (cf. 5.5.6).

The FFT routine given below is based on one originally written by N. M. Brenner. The input quantities are the number of complex data points (nn), the data array (data), and isign, which should be set to either ± 1 and is the sign of i in the exponential of equation (12.1.7). When isign is set to -1, the routine thus calculates the inverse transform (12.1.9) — except that it does not multiply by the normalizing factor $1/N$ that appears in that equation. You can do that yourself.

Notice that the argument nn is the number of *complex* data points, although

we avoid the use of complex arithmetic because of the inefficient implementations found on many computers. The actual length of the real array (data) is 2 times nn, with each complex value occupying two consecutive locations. In other words, data(1) is the real part of f_0, data(2) is the imaginary part of f_0, and so on up to data(2*nn-1), which is the real part of f_{N-1}, and data(2*nn), which is the imaginary part of f_{N-1}. The FFT routine returns the F_n's packed in exactly the same fashion, as nn complex numbers.

The real and imaginary parts of the zero frequency component F_0 are in data(1) and data(2); the smallest nonzero positive frequency has real and imaginary parts in data(3) and data(4); the smallest (in magnitude) nonzero negative frequency has real and imaginary parts in data(2*nn-1) and data(2*nn). Positive frequencies increasing in magnitude are stored in the real-imaginary pairs data(5), data(6) up to data(nn-1), data(nn). Negative frequencies of increasing magnitude are stored in data(2*nn-3), data(2*nn-2) down to data(nn+3), data(nn+4). Finally, the pair data(nn+1), data(nn+2) contain the real and imaginary parts of the one aliased point that contains the most positive and the most negative frequency. You should try to develop a familiarity with this storage arrangement of complex spectra, also shown in Figure 12.2.2, since it is the practical standard.

```
      SUBROUTINE four1(data,nn,isign)
      INTEGER isign,nn
      REAL data(2*nn)
         Replaces data(1:2*nn) by its discrete Fourier transform, if isign is input as 1; or replaces
         data(1:2*nn) by nn times its inverse discrete Fourier transform, if isign is input as -1.
         data is a complex array of length nn or, equivalently, a real array of length 2*nn. nn
         MUST be an integer power of 2 (this is not checked for!).
      INTEGER i,istep,j,m,mmax,n
      REAL tempi,tempr
      DOUBLE PRECISION theta,wi,wpi,wpr,wr,wtemp    Double precision for the trigonomet-
      n=2*nn                                        ric recurrences.
      j=1
      do 11 i=1,n,2                      This is the bit-reversal section of the routine.
        if(j.gt.i)then
          tempr=data(j)                  Exchange the two complex numbers.
          tempi=data(j+1)
          data(j)=data(i)
          data(j+1)=data(i+1)
          data(i)=tempr
          data(i+1)=tempi
        endif
        m=n/2
1       if ((m.ge.2).and.(j.gt.m)) then
          j=j-m
          m=m/2
        goto 1
        endif
        j=j+m
      enddo 11
      mmax=2                             Here begins the Danielson-Lanczos section of the routine.
2     if (n.gt.mmax) then               Outer loop executed log_2 nn times.
        istep=2*mmax
        theta=6.28318530717959d0/(isign*mmax)        Initialize for the trigonometric recur-
        wpr=-2.d0*sin(0.5d0*theta)**2                rence.
        wpi=sin(theta)
        wr=1.d0
        wi=0.d0
        do 13 m=1,mmax,2              Here are the two nested inner loops.
```

Figure 12.2.2.   Input and output arrays for FFT. (a) The input array contains N (a power of 2) complex time samples in a real array of length $2N$, with real and imaginary parts alternating. (b) The output array contains the complex Fourier spectrum at N values of frequency. Real and imaginary parts again alternate. The array starts with zero frequency, works up to the most positive frequency (which is ambiguous with the most negative frequency). Negative frequencies follow, from the second-most negative up to the frequency just below zero.

```
      do 12 i=m,n,istep
          j=i+mmax              This is the Danielson-Lanczos formula:
          tempr=sngl(wr)*data(j)-sngl(wi)*data(j+1)
          tempi=sngl(wr)*data(j+1)+sngl(wi)*data(j)
          data(j)=data(i)-tempr
          data(j+1)=data(i+1)-tempi
          data(i)=data(i)+tempr
          data(i+1)=data(i+1)+tempi
      enddo 12
      wtemp=wr                  Trigonometric recurrence.
      wr=wr*wpr-wi*wpi+wr
      wi=wi*wpr+wtemp*wpi+wi
    enddo 13
    mmax=istep
goto 2                          Not yet done.
endif                           All done.
return
END
```

(A double precision version of four1, named dfour1, is used by the routine mpmul in §20.6. You can easily make the conversion, or else get the converted routine from the *Numerical Recipes* diskette.)

Other FFT Algorithms

We should mention that there are a number of variants on the basic FFT algorithm given above. As we have seen, that algorithm first rearranges the input elements into bit-reverse order, then builds up the output transform in $\log_2 N$ iterations. In the literature, this sequence is called a *decimation-in-time* or *Cooley-Tukey* FFT algorithm. It is also possible to derive FFT algorithms that first go through a set of $\log_2 N$ iterations on the input data, and rearrange the *output* values into bit-reverse order. These are called *decimation-in-frequency* or *Sande-Tukey* FFT algorithms. For some applications, such as convolution (§13.1), one takes a data set into the Fourier domain and then, after some manipulation, back out again. In these cases it is possible to avoid all bit reversing. You use a decimation-in-frequency algorithm (without its bit reversing) to get into the "scrambled" Fourier domain, do your operations there, and then use an inverse algorithm (without *its* bit reversing) to get back to the time domain. While elegant in principle, this procedure does not in practice save much computation time, since the bit reversals represent only a small fraction of an FFT's operations count, and since most useful operations in the frequency domain require a knowledge of which points correspond to which frequencies.

Another class of FFTs subdivides the initial data set of length N not all the way down to the trivial transform of length 1, but rather only down to some other small power of 2, for example $N = 4$, *base-4 FFTs*, or $N = 8$, *base-8 FFTs*. These small transforms are then done by small sections of highly optimized coding which take advantage of special symmetries of that particular small N. For example, for $N = 4$, the trigonometric sines and cosines that enter are all ± 1 or 0, so many multiplications are eliminated, leaving largely additions and subtractions. These can be faster than simpler FFTs by some significant, but not overwhelming, factor, e.g., 20 or 30 percent.

There are also FFT algorithms for data sets of length N not a power of two. They work by using relations analogous to the Danielson-Lanczos Lemma to subdivide the initial problem into successively smaller problems, not by factors of 2, but by whatever small prime factors happen to divide N. The larger that the largest prime factor of N is, the worse this method works. If N is prime, then no subdivision is possible, and the user (whether he knows it or not) is taking a *slow* Fourier transform, of order N^2 instead of order $N \log_2 N$. Our advice is to stay clear of such FFT implementations, with perhaps one class of exceptions, the *Winograd Fourier transform algorithms*. Winograd algorithms are in some ways analogous to the base-4 and base-8 FFTs. Winograd has derived highly optimized codings for taking small-N discrete Fourier transforms, e.g., for $N = 2, 3, 4, 5, 7, 8, 11, 13, 16$. The algorithms also use a new and clever way of combining the subfactors. The method involves a reordering of the data both before the hierarchical processing and after it, but it allows a significant reduction in the number of multiplications in the algorithm. For some especially favorable values of N, the Winograd algorithms can be significantly (e.g., up to a factor of 2) faster than the simpler FFT algorithms of the nearest integer power of 2. This advantage in speed, however, must be weighed against the considerably more complicated data indexing involved in these transforms, and the fact that the Winograd transform cannot be done "in place."

Finally, an interesting class of transforms for doing convolutions quickly are number theoretic transforms. These schemes replace floating-point arithmetic with

integer arithmetic modulo some large prime $N+1$, and the Nth root of 1 by the modulo arithmetic equivalent. Strictly speaking, these are not *Fourier* transforms at all, but the properties are quite similar and computational speed can be far superior. On the other hand, their use is somewhat restricted to quantities like correlations and convolutions since the transform itself is not easily interpretable as a "frequency" spectrum.

CITED REFERENCES AND FURTHER READING:

Nussbaumer, H.J. 1982, *Fast Fourier Transform and Convolution Algorithms* (New York: Springer-Verlag).

Elliott, D.F., and Rao, K.R. 1982, *Fast Transforms: Algorithms, Analyses, Applications* (New York: Academic Press).

Brigham, E.O. 1974, *The Fast Fourier Transform* (Englewood Cliffs, NJ: Prentice-Hall). [1]

Bloomfield, P. 1976, *Fourier Analysis of Time Series – An Introduction* (New York: Wiley).

Van Loan, C. 1992, *Computational Frameworks for the Fast Fourier Transform* (Philadelphia: S.I.A.M.).

Beauchamp, K.G. 1984, *Applications of Walsh Functions and Related Functions* (New York: Academic Press) [non-Fourier transforms].

Heideman, M.T., Johnson, D.H., and Burris, C.S. 1984, *IEEE ASSP Magazine*, pp. 14–21 (October).

12.3 FFT of Real Functions, Sine and Cosine Transforms

It happens frequently that the data whose FFT is desired consist of real-valued samples f_j, $j = 0 \ldots N - 1$. To use four1, we put these into a complex array with all imaginary parts set to zero. The resulting transform F_n, $n = 0 \ldots N - 1$ satisfies $F_{N-n}{}^* = F_n$. Since this complex-valued array has real values for F_0 and $F_{N/2}$, and $(N/2) - 1$ other independent values $F_1 \ldots F_{N/2-1}$, it has the same $2(N/2 - 1) + 2 = N$ "degrees of freedom" as the original, real data set. However, the use of the full complex FFT algorithm for real data is inefficient, both in execution time and in storage required. You would think that there is a better way.

There are *two* better ways. The first is "mass production": Pack two separate real functions into the input array in such a way that their individual transforms can be separated from the result. This is implemented in the program twofft below. This may remind you of a one-cent sale, at which you are coerced to purchase two of an item when you only need one. However, remember that for correlations and convolutions the Fourier transforms of two functions are involved, and this is a handy way to do them both at once. The second method is to pack the real input array cleverly, without extra zeros, into a complex array of half its length. One then performs a complex FFT on this shorter length; the trick is then to get the required answer out of the result. This is done in the program realft below.

Transform of Two Real Functions Simultaneously

First we show how to exploit the symmetry of the transform F_n to handle two real functions at once: Since the input data f_j are real, the components of the discrete Fourier transform satisfy

$$F_{N-n} = (F_n)^* \qquad (12.3.1)$$

where the asterisk denotes complex conjugation. By the same token, the discrete Fourier transform of a purely imaginary set of g_j's has the opposite symmetry.

$$G_{N-n} = -(G_n)^* \qquad (12.3.2)$$

Therefore we can take the discrete Fourier transform of two real functions each of length N simultaneously by packing the two data arrays as the real and imaginary parts, respectively, of the complex input array of four1. Then the resulting transform array can be unpacked into two complex arrays with the aid of the two symmetries. Routine twofft works out these ideas.

```
      SUBROUTINE twofft(data1,data2,fft1,fft2,n)
      INTEGER n
      REAL data1(n),data2(n)
      COMPLEX fft1(n),fft2(n)
C     USES four1
      Given two real input arrays data1(1:n) and data2(1:n), this routine calls four1 and
      returns two complex output arrays, fft1(1:n) and fft2(1:n), each of complex length n
      (i.e., real length 2*n), which contain the discrete Fourier transforms of the respective data
      arrays. n MUST be an integer power of 2.
      INTEGER j,n2
      COMPLEX h1,h2,c1,c2
      c1=cmplx(0.5,0.0)
      c2=cmplx(0.0,-0.5)
      do 11 j=1,n
          fft1(j)=cmplx(data1(j),data2(j))    Pack the two real arrays into one complex
      enddo 11                                  array.
      call four1(fft1,n,1)                    Transform the complex array.
      fft2(1)=cmplx(aimag(fft1(1)),0.0)
      fft1(1)=cmplx(real(fft1(1)),0.0)
      n2=n+2
      do 12 j=2,n/2+1
          h1=c1*(fft1(j)+conjg(fft1(n2-j)))   Use symmetries to separate the two trans-
          h2=c2*(fft1(j)-conjg(fft1(n2-j)))     forms.
          fft1(j)=h1                          Ship them out in two complex arrays.
          fft1(n2-j)=conjg(h1)
          fft2(j)=h2
          fft2(n2-j)=conjg(h2)
      enddo 12
      return
      END
```

What about the reverse process? Suppose you have two complex transform arrays, each of which has the symmetry (12.3.1), so that you know that the inverses of both transforms are real functions. Can you invert both in a single FFT? This is even easier than the other direction. Use the fact that the FFT is linear and form the sum of the first transform plus i times the second. Invert using four1 with

$\mathtt{isign} = -1$. The real and imaginary parts of the resulting complex array are the two desired real functions.

FFT of Single Real Function

To implement the second method, which allows us to perform the FFT of a *single* real function without redundancy, we split the data set in half, thereby forming two real arrays of half the size. We can apply the program above to these two, but of course the result will not be the transform of the original data. It will be a schizophrenic combination of two transforms, each of which has half of the information we need. Fortunately, this schizophrenia is treatable. It works like this:

The right way to split the original data is to take the even-numbered f_j as one data set, and the odd-numbered f_j as the other. The beauty of this is that we can take the original real array and treat it as a complex array h_j of half the length. The first data set is the real part of this array, and the second is the imaginary part, as prescribed for $\mathtt{twofft}$. No repacking is required. In other words $h_j = f_{2j} + if_{2j+1}, \quad j = 0, \ldots, N/2 - 1$. We submit this to $\mathtt{four1}$, and it will return a complex array $H_n = F_n^e + iF_n^o, \quad n = 0, \ldots, N/2 - 1$ with

$$
\begin{aligned}
F_n^e &= \sum_{k=0}^{N/2-1} f_{2k}\, e^{2\pi ikn/(N/2)} \\
F_n^o &= \sum_{k=0}^{N/2-1} f_{2k+1}\, e^{2\pi ikn/(N/2)}
\end{aligned}
\tag{12.3.3}
$$

The discussion of program $\mathtt{twofft}$ tells you how to separate the two transforms F_n^e and F_n^o out of H_n. How do you work them into the transform F_n of the original data set f_j? Simply glance back at equation (12.2.3):

$$
F_n = F_n^e + e^{2\pi in/N} F_n^o \qquad n = 0, \ldots, N - 1
\tag{12.3.4}
$$

Expressed directly in terms of the transform H_n of our real (masquerading as complex) data set, the result is

$$
F_n = \frac{1}{2}(H_n + H_{N/2-n}{}^*) - \frac{i}{2}(H_n - H_{N/2-n}{}^*)e^{2\pi in/N} \qquad n = 0, \ldots, N - 1
\tag{12.3.5}
$$

A few remarks:
- Since $F_{N-n}{}^* = F_n$ there is no point in saving the entire spectrum. The positive frequency half is sufficient and can be stored in the same array as the original data. The operation can, in fact, be done in place.
- Even so, we need values $H_n, \ n = 0, \ldots, N/2$ whereas $\mathtt{four1}$ returns only the values $n = 0, \ldots, N/2 - 1$. Symmetry to the rescue, $H_{N/2} = H_0$.
- The values F_0 and $F_{N/2}$ are real and independent. In order to actually get the entire F_n in the original array space, it is convenient to return $F_{N/2}$ as the imaginary part of F_0.

- Despite its complicated form, the process above is invertible. First peel $F_{N/2}$ out of F_0. Then construct

$$F_n^e = \frac{1}{2}(F_n + F_{N/2-n}^*)$$
$$F_n^o = \frac{1}{2}e^{-2\pi in/N}(F_n - F_{N/2-n}^*)$$
$$n = 0, \ldots, N/2 - 1 \quad (12.3.6)$$

and use `four1` to find the inverse transform of $H_n = F_n^{(1)} + iF_n^{(2)}$. Surprisingly, the actual algebraic steps are virtually identical to those of the forward transform.

Here is a representation of what we have said:

```
      SUBROUTINE realft(data,n,isign)
      INTEGER isign,n
      REAL data(n)
C     USES four1
          Calculates the Fourier transform of a set of n real-valued data points. Replaces this data
          (which is stored in array data(1:n)) by the positive frequency half of its complex Fourier
          transform. The real-valued first and last components of the complex transform are returned
          as elements data(1) and data(2), respectively. n must be a power of 2. This routine
          also calculates the inverse transform of a complex data array if it is the transform of real
          data. (Result in this case must be multiplied by 2/n.)
      INTEGER i,i1,i2,i3,i4,n2p3
      REAL c1,c2,h1i,h1r,h2i,h2r,wis,wrs
      DOUBLE PRECISION theta,wi,wpi,wpr,
     *     wr,wtemp                      Double precision for the trigonometric recurrences.
      theta=3.141592653589793d0/dble(n/2) Initialize the recurrence.
      c1=0.5
      if (isign.eq.1) then
          c2=-0.5
          call four1(data,n/2,+1)        The forward transform is here.
      else
          c2=0.5                         Otherwise set up for an inverse transform.
          theta=-theta
      endif
      wpr=-2.0d0*sin(0.5d0*theta)**2
      wpi=sin(theta)
      wr=1.0d0+wpr
      wi=wpi
      n2p3=n+3
      do 11 i=2,n/4                      Case i=1 done separately below.
          i1=2*i-1
          i2=i1+1
          i3=n2p3-i2
          i4=i3+1
          wrs=sngl(wr)
          wis=sngl(wi)
          h1r=c1*(data(i1)+data(i3))     The two separate transforms are separated out of
          h1i=c1*(data(i2)-data(i4))         data.
          h2r=-c2*(data(i2)+data(i4))
          h2i=c2*(data(i1)-data(i3))
          data(i1)=h1r+wrs*h2r-wis*h2i   Here they are recombined to form the true trans-
          data(i2)=h1i+wrs*h2i+wis*h2r       form of the original real data.
          data(i3)=h1r-wrs*h2r+wis*h2i
          data(i4)=-h1i+wrs*h2i+wis*h2r
          wtemp=wr                       The recurrence.
          wr=wr*wpr-wi*wpi+wr
          wi=wi*wpr+wtemp*wpi+wi
      enddo 11
```

```
if (isign.eq.1) then
    h1r=data(1)
    data(1)=h1r+data(2)
    data(2)=h1r-data(2)              Squeeze the first and last data together to get
else                                     them all within the original array.
    h1r=data(1)
    data(1)=c1*(h1r+data(2))
    data(2)=c1*(h1r-data(2))
    call four1(data,n/2,-1)          This is the inverse transform for the case isign=-1.
endif
return
END
```

Fast Sine and Cosine Transforms

Among their other uses, the Fourier transforms of functions can be used to solve differential equations (see §19.4). The most common boundary conditions for the solutions are 1) they have the value zero at the boundaries, or 2) their derivatives are zero at the boundaries. In the first instance, the natural transform to use is the *sine* transform, given by

$$F_k = \sum_{j=1}^{N-1} f_j \sin(\pi jk/N) \qquad \text{sine transform} \qquad (12.3.7)$$

where f_j, $j = 0, \ldots, N - 1$ is the data array, and $f_0 \equiv 0$.

At first blush this appears to be simply the imaginary part of the discrete Fourier transform. However, the argument of the sine differs by a factor of two from the value that would make this so. The sine transform uses *sines only* as a complete set of functions in the interval from 0 to 2π, and, as we shall see, the cosine transform uses *cosines only*. By contrast, the normal FFT uses both sines and cosines, but only half as many of each. (See Figure 12.3.1.)

The expression (12.3.7) can be "force-fit" into a form that allows its calculation via the FFT. The idea is to extend the given function rightward past its last tabulated value. We extend the data to twice their length in such a way as to make them an *odd* function about $j = N$, with $f_N = 0$,

$$f_{2N-j} \equiv -f_j \qquad j = 0, \ldots, N - 1 \qquad (12.3.8)$$

Consider the FFT of this extended function:

$$F_k = \sum_{j=0}^{2N-1} f_j e^{2\pi ijk/(2N)} \qquad (12.3.9)$$

The half of this sum from $j = N$ to $j = 2N - 1$ can be rewritten with the substitution $j' = 2N - j$

$$\sum_{j=N}^{2N-1} f_j e^{2\pi ijk/(2N)} = \sum_{j'=1}^{N} f_{2N-j'} e^{2\pi i(2N-j')k/(2N)}$$

$$= -\sum_{j'=0}^{N-1} f_{j'} e^{-2\pi ij'k/(2N)} \qquad (12.3.10)$$
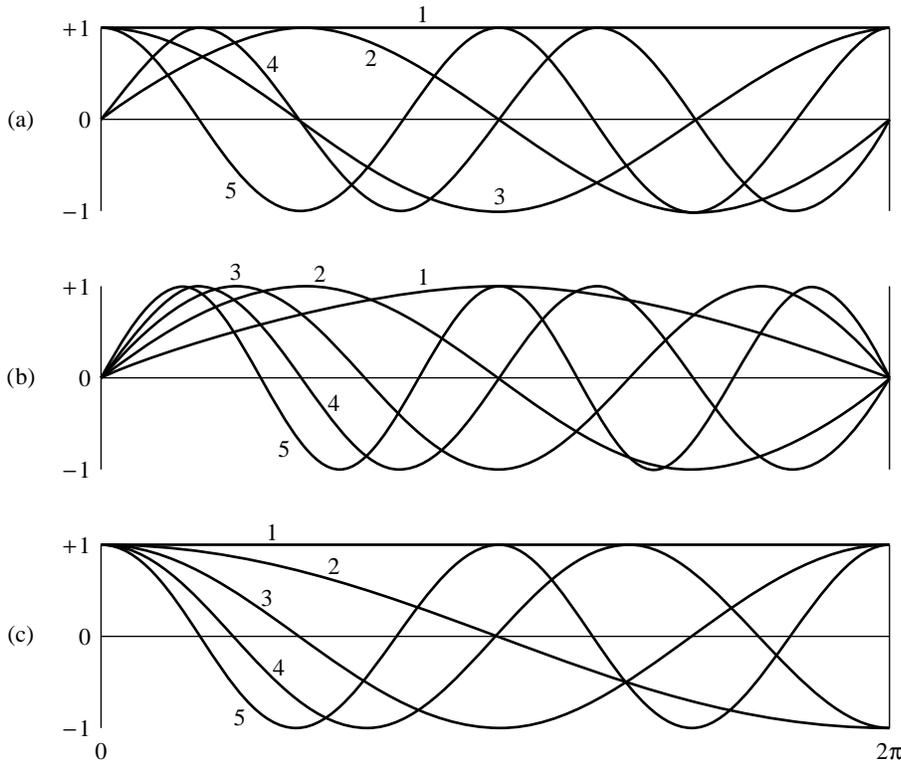
Figure 12.3.1. Basis functions used by the Fourier transform (a), sine transform (b), and cosine transform (c), are plotted. The first five basis functions are shown in each case. (For the Fourier transform, the real and imaginary parts of the basis functions are both shown.) While some basis functions occur in more than one transform, the basis sets are distinct. For example, the sine transform functions labeled (1), (3), (5) are not present in the Fourier basis. Any of the three sets can expand any function in the interval shown; however, the sine or cosine transform best expands functions matching the boundary conditions of the respective basis functions, namely zero function values for sine, zero derivatives for cosine.

so that

$$F_k = \sum_{j=0}^{N-1} f_j \left[e^{2\pi ijk/(2N)} - e^{-2\pi ijk/(2N)} \right]$$

$$= 2i \sum_{j=0}^{N-1} f_j \sin(\pi jk/N)$$

(12.3.11)

Thus, up to a factor $2i$ we get the sine transform from the FFT of the extended function.

This method introduces a factor of two inefficiency into the computation by extending the data. This inefficiency shows up in the FFT output, which has zeros for the real part of every element of the transform. For a one-dimensional problem, the factor of two may be bearable, especially in view of the simplicity of the method. When we work with partial differential equations in two or three dimensions, though, the factor becomes four or eight, so efforts to eliminate the inefficiency are well rewarded.

From the original real data array f_j we will construct an auxiliary array y_j and apply to it the routine realft. The output will then be used to construct the desired transform. For the sine transform of data f_j, $j = 1, \ldots, N-1$, the auxiliary array is

$$y_0 = 0$$
$$y_j = \sin(j\pi/N)(f_j + f_{N-j}) + \frac{1}{2}(f_j - f_{N-j}) \qquad j = 1, \ldots, N-1 \tag{12.3.12}$$

This array is of the same dimension as the original. Notice that the first term is symmetric about $j = N/2$ and the second is antisymmetric. Consequently, when realft is applied to y_j, the result has real parts R_k and imaginary parts I_k given by

$$
\begin{aligned}
R_k &= \sum_{j=0}^{N-1} y_j \cos(2\pi jk/N) \\
&= \sum_{j=1}^{N-1} (f_j + f_{N-j})\sin(j\pi/N)\cos(2\pi jk/N) \\
&= \sum_{j=0}^{N-1} 2f_j \sin(j\pi/N)\cos(2\pi jk/N) \\
&= \sum_{j=0}^{N-1} f_j \left[\sin\frac{(2k+1)j\pi}{N} - \sin\frac{(2k-1)j\pi}{N} \right] \\
&= F_{2k+1} - F_{2k-1}
\end{aligned}
\tag{12.3.13}
$$

$$
\begin{aligned}
I_k &= \sum_{j=0}^{N-1} y_j \sin(2\pi jk/N) \\
&= \sum_{j=1}^{N-1} (f_j - f_{N-j})\frac{1}{2}\sin(2\pi jk/N) \\
&= \sum_{j=0}^{N-1} f_j \sin(2\pi jk/N) \\
&= F_{2k}
\end{aligned}
\tag{12.3.14}
$$

Therefore F_k can be determined as follows:

$$F_{2k} = I_k \qquad F_{2k+1} = F_{2k-1} + R_k \qquad k = 0, \ldots, (N/2-1) \tag{12.3.15}$$

The even terms of F_k are thus determined very directly. The odd terms require a recursion, the starting point of which follows from setting $k = 0$ in equation (12.3.15) and using $F_1 = -F_{-1}$:

$$F_1 = \frac{1}{2}R_0 \tag{12.3.16}$$

The implementing program is

```
      SUBROUTINE sinft(y,n)
      INTEGER n
      REAL y(n)
C   USES realft
      Calculates the sine transform of a set of n real-valued data points stored in array y(1:n).
      The number n must be a power of 2. On exit y is replaced by its transform. This program,
      without changes, also calculates the inverse sine transform, but in this case the output array
      should be multiplied by 2/n.
      INTEGER j
      REAL sum,y1,y2
      DOUBLE PRECISION theta,wi,wpi,wpr,
     *    wr,wtemp                        Double precision in the trigonometric recurrences.
      theta=3.141592653589793d0/dble(n)  Initialize the recurrence.
      wr=1.0d0
      wi=0.0d0
      wpr=-2.0d0*sin(0.5d0*theta)**2
      wpi=sin(theta)
      y(1)=0.0
      do 11 j=1,n/2
          wtemp=wr
          wr=wr*wpr-wi*wpi+wr             Calculate the sine for the auxiliary array.
          wi=wi*wpr+wtemp*wpi+wi          The cosine is needed to continue the recurrence.
          y1=wi*(y(j+1)+y(n-j+1))         Construct the auxiliary array.
          y2=0.5*(y(j+1)-y(n-j+1))
          y(j+1)=y1+y2                    Terms j and N − j are related.
          y(n-j+1)=y1-y2
      enddo 11
      call realft(y,n,+1)                 Transform the auxiliary array.
      sum=0.0
      y(1)=0.5*y(1)                       Initialize the sum used for odd terms below.
      y(2)=0.0
      do 12 j=1,n-1,2
          sum=sum+y(j)
          y(j)=y(j+1)                     Even terms in the transform are determined directly.
          y(j+1)=sum                      Odd terms are determined by this running sum.
      enddo 12
      return
      END
```

The sine transform, curiously, is its own inverse. If you apply it twice, you get the original data, but multiplied by a factor of $N/2$.

The other common boundary condition for differential equations is that the derivative of the function is zero at the boundary. In this case the natural transform is the *cosine* transform. There are several possible ways of defining the transform. Each can be thought of as resulting from a different way of extending a given array to create an even array of double the length, and/or from whether the extended array contains $2N - 1$, $2N$, or some other number of points. In practice, only two of the numerous possibilities are useful so we will restrict ourselves to just these two.

The first form of the cosine transform uses $N + 1$ data points:

$$F_k = \frac{1}{2}[f_0 + (-1)^k f_N] + \sum_{j=1}^{N-1} f_j \cos(\pi jk/N) \qquad (12.3.17)$$

It results from extending the given array to an even array about $j = N$, with

$$f_{2N-j} = f_j, \qquad j = 0, \ldots, N - 1 \qquad (12.3.18)$$

If you substitute this extended array into equation (12.3.9), and follow steps analogous to those leading up to equation (12.3.11), you will find that the Fourier transform is just twice the cosine transform (12.3.17). Another way of thinking about the formula (12.3.17) is to notice that it is the Chebyshev Gauss-Lobatto quadrature formula (see §4.5), often used in Clenshaw-Curtis adaptive quadrature (see §5.9, equation 5.9.4).

Once again the transform can be computed without the factor of two inefficiency. In this case the auxiliary function is

$$y_j = \frac{1}{2}(f_j + f_{N-j}) - \sin(j\pi/N)(f_j - f_{N-j}) \qquad j = 0, \ldots, N-1 \quad (12.3.19)$$

Instead of equation (12.3.15), `realft` now gives

$$F_{2k} = R_k \qquad F_{2k+1} = F_{2k-1} + I_k \qquad k = 0, \ldots, (N/2 - 1) \quad (12.3.20)$$

The starting value for the recursion for odd k in this case is

$$F_1 = \frac{1}{2}(f_0 - f_N) + \sum_{j=1}^{N-1} f_j \cos(j\pi/N) \quad (12.3.21)$$

This sum does not appear naturally among the R_k and I_k, and so we accumulate it during the generation of the array y_j.

Once again this transform is its own inverse, and so the following routine works for both directions of the transformation. Note that although this form of the cosine transform has $N + 1$ input and output values, it passes an array only of length N to `realft`.

```
      SUBROUTINE cosft1(y,n)
      INTEGER n
      REAL y(n+1)
C     USES realft
        Calculates the cosine transform of a set y(1:n+1) of real-valued data points. The trans-
        formed data replace the original data in array y. n must be a power of 2. This program,
        without changes, also calculates the inverse cosine transform, but in this case the output
        array should be multiplied by 2/n.
      INTEGER j
      REAL sum,y1,y2
      DOUBLE PRECISION theta,wi,wpi,wpr,wr,wtemp        For trig. recurrences.
      theta=3.141592653589793d0/n            Initialize the recurrence.
      wr=1.0d0
      wi=0.0d0
      wpr=-2.0d0*sin(0.5d0*theta)**2
      wpi=sin(theta)
      sum=0.5*(y(1)-y(n+1))
      y(1)=0.5*(y(1)+y(n+1))
      do 11 j=1,n/2-1                        j=n/2 unnecessary since y(n/2+1) unchanged.
        wtemp=wr
        wr=wr*wpr-wi*wpi+wr                  Carry out the recurrence.
        wi=wi*wpr+wtemp*wpi+wi
        y1=0.5*(y(j+1)+y(n-j+1))            Calculate the auxiliary function.
        y2=(y(j+1)-y(n-j+1))
        y(j+1)=y1-wi*y2                     The values for j and N − j are related.
        y(n-j+1)=y1+wi*y2
        sum=sum+wr*y2                        Carry along this sum for later use in unfolding the
      enddo 11                                   transform.
```

```
call realft(y,n,+1)                    Calculate the transform of the auxiliary function.
y(n+1)=y(2)
y(2)=sum                               sum is the value of F₁ in equation (12.3.21).
do 12 j=4,n,2
    sum=sum+y(j)                       Equation (12.3.20).
    y(j)=sum
enddo 12
return
END
```

The second important form of the cosine transform is defined by

$$F_k = \sum_{j=0}^{N-1} f_j \cos \frac{\pi k(j + \frac{1}{2})}{N} \qquad (12.3.22)$$

with inverse

$$f_j = \frac{2}{N} \sum_{k=0}^{N-1}{}' F_k \cos \frac{\pi k(j + \frac{1}{2})}{N} \qquad (12.3.23)$$

Here the prime on the summation symbol means that the term for $k = 0$ has a coefficient of $\frac{1}{2}$ in front. This form arises by extending the given data, defined for $j = 0, \ldots, N-1$, to $j = N, \ldots, 2N-1$ in such a way that it is even about the point $N - \frac{1}{2}$ and periodic. (It is therefore also even about $j = -\frac{1}{2}$.) The form (12.3.23) is related to Gauss-Chebyshev quadrature (see equation 4.5.19), to Chebyshev approximation (§5.8, equation 5.8.7), and Clenshaw-Curtis quadrature (§5.9).

This form of the cosine transform is useful when solving differential equations on "staggered" grids, where the variables are centered midway between mesh points. It is also the standard form in the field of data compression and image processing.

The auxiliary function used in this case is similar to equation (12.3.19):

$$y_j = \frac{1}{2}(f_j + f_{N-j-1}) - \sin \frac{\pi(j + \frac{1}{2})}{N}(f_j - f_{N-j-1}) \qquad j = 0, \ldots, N-1 \tag{12.3.24}$$

Carrying out the steps similar to those used to get from (12.3.12) to (12.3.15), we find

$$F_{2k} = \cos \frac{\pi k}{N} R_k - \sin \frac{\pi k}{N} I_k \qquad (12.3.25)$$

$$F_{2k-1} = \sin \frac{\pi k}{N} R_k + \cos \frac{\pi k}{N} I_k + F_{2k+1} \qquad (12.3.26)$$

Note that equation (12.3.26) gives

$$F_{N-1} = \frac{1}{2} R_{N/2} \qquad (12.3.27)$$

Thus the even components are found directly from (12.3.25), while the odd components are found by recursing (12.3.26) down from $k = N/2 - 1$, using (12.3.27) to start.

Since the transform is not self-inverting, we have to reverse the above steps to find the inverse. Here is the routine:

```
      SUBROUTINE cosft2(y,n,isign)
      INTEGER isign,n
      REAL y(n)
C     USES realft
      Calculates the "staggered" cosine transform of a set y(1:n) of real-valued data points.
      The transformed data replace the original data in array y. n must be a power of 2. Set
      isign to +1 for a transform, and to −1 for an inverse transform. For an inverse transform,
      the output array should be multiplied by 2/n.
      INTEGER i
      REAL sum,sum1,y1,y2,ytemp
      DOUBLE PRECISION theta,wi,wi1,wpi,wpr,wr,wr1,wtemp,PI
         Double precision for the trigonometric recurrences.
      PARAMETER (PI=3.141592653589793d0)
      theta=0.5d0*PI/n                    Initialize the recurrences.
      wr=1.0d0
      wi=0.0d0
      wr1=cos(theta)
      wi1=sin(theta)
      wpr=-2.0d0*wi1**2
      wpi=sin(2.d0*theta)
      if(isign.eq.1)then                  Forward transform.
         do 11 i=1,n/2
            y1=0.5*(y(i)+y(n-i+1))        Calculate the auxiliary function.
            y2=wi1*(y(i)-y(n-i+1))
            y(i)=y1+y2
            y(n-i+1)=y1-y2
            wtemp=wr1                      Carry out the recurrence.
            wr1=wr1*wpr-wi1*wpi+wr1
            wi1=wi1*wpr+wtemp*wpi+wi1
         enddo 11
         call realft(y,n,1)               Calculate the transform of the auxiliary function.
         do 12 i=3,n,2                     Even terms.
            wtemp=wr
            wr=wr*wpr-wi*wpi+wr
            wi=wi*wpr+wtemp*wpi+wi
            y1=y(i)*wr-y(i+1)*wi
            y2=y(i+1)*wr+y(i)*wi
            y(i)=y1
            y(i+1)=y2
         enddo 12
         sum=0.5*y(2)                      Initialize recurrence for odd terms with $\frac{1}{2}R_{N/2}$.
         do 13 i=n,2,-2                     Carry out recurrence for odd terms.
            sum1=sum
            sum=sum+y(i)
            y(i)=sum1
         enddo 13
      else if(isign.eq.-1)then            Inverse transform.
         ytemp=y(n)
         do 14 i=n,4,-2                    Form difference of odd terms.
            y(i)=y(i-2)-y(i)
         enddo 14
         y(2)=2.0*ytemp
         do 15 i=3,n,2                     Calculate $R_k$ and $I_k$.
            wtemp=wr
            wr=wr*wpr-wi*wpi+wr
            wi=wi*wpr+wtemp*wpi+wi
            y1=y(i)*wr+y(i+1)*wi
            y2=y(i+1)*wr-y(i)*wi
            y(i)=y1
            y(i+1)=y2
         enddo 15
         call realft(y,n,-1)
         do 16 i=1,n/2                     Invert auxiliary array.
            y1=y(i)+y(n-i+1)
```

```
        y2=(0.5/wi1)*(y(i)-y(n-i+1))
        y(i)=0.5*(y1+y2)
        y(n-i+1)=0.5*(y1-y2)
        wtemp=wr1
        wr1=wr1*wpr-wi1*wpi+wr1
        wi1=wi1*wpr+wtemp*wpi+wi1
    enddo 16
endif
return
END
```

An alternative way of implementing this algorithm is to form an auxiliary function by copying the even elements of f_j into the first $N/2$ locations, and the odd elements into the next $N/2$ elements in reverse order. However, it is not easy to implement the alternative algorithm without a temporary storage array and we prefer the above in-place algorithm.

Finally, we mention that there exist fast cosine transforms for small N that do not rely on an auxiliary function or use an FFT routine. Instead, they carry out the transform directly, often coded in hardware for fixed N of small dimension [1].

CITED REFERENCES AND FURTHER READING:

Brigham, E.O. 1974, *The Fast Fourier Transform* (Englewood Cliffs, NJ: Prentice-Hall), §10–10.

Sorensen, H.V., Jones, D.L., Heideman, M.T., and Burris, C.S. 1987, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, pp. 849–863.

Hou, H.S. 1987, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, pp. 1455–1461 [see for additional references].

Hockney, R.W. 1971, in *Methods in Computational Physics*, vol. 9 (New York: Academic Press).

Temperton, C. 1980, *Journal of Computational Physics*, vol. 34, pp. 314–329.

Clarke, R.J. 1985, *Transform Coding of Images*, (Reading, MA: Addison-Wesley).

Gonzalez, R.C., and Wintz, P. 1987, *Digital Image Processing*, (Reading, MA: Addison-Wesley).

Chen, W., Smith, C.H., and Fralick, S.C. 1977, *IEEE Transactions on Communications*, vol. COM-25, pp. 1004–1009. [1]

12.4 FFT in Two or More Dimensions

Given a complex function $h(k_1, k_2)$ defined over the two-dimensional grid $0 \le k_1 \le N_1 - 1$, $0 \le k_2 \le N_2 - 1$, we can define its two-dimensional discrete Fourier transform as a complex function $H(n_1, n_2)$, defined over the same grid,

$$H(n_1, n_2) \equiv \sum_{k_2=0}^{N_2-1} \sum_{k_1=0}^{N_1-1} \exp(2\pi i k_2 n_2 / N_2) \, \exp(2\pi i k_1 n_1 / N_1) \, h(k_1, k_2)$$

(12.4.1)

By pulling the "subscripts 2" exponential outside of the sum over k_1, or by reversing the order of summation and pulling the "subscripts 1" outside of the sum over k_2,

we can see instantly that the two-dimensional FFT can be computed by taking one-dimensional FFTs sequentially on each index of the original function. Symbolically,

$$
\begin{aligned}
H(n_1, n_2) &= \text{FFT-on-index-1}\,(\text{FFT-on-index-2}\,[h(k_1, k_2)]) \\
&= \text{FFT-on-index-2}\,(\text{FFT-on-index-1}\,[h(k_1, k_2)])
\end{aligned}
\tag{12.4.2}
$$

For this to be practical, of course, both N_1 and N_2 should be some efficient length for an FFT, usually a power of 2. Programming a two-dimensional FFT, using (12.4.2) with a one-dimensional FFT routine, is a bit clumsier than it seems at first. Because the one-dimensional routine requires that its input be in consecutive order as a one-dimensional complex array, you find that you are endlessly copying things out of the multidimensional input array and then copying things back into it. This is not recommended technique. Rather, you should use a multidimensional FFT routine, such as the one we give below.

The generalization of (12.4.1) to more than two dimensions, say to L-dimensions, is evidently

$$
H(n_1, \ldots, n_L) \equiv \sum_{k_L=0}^{N_L-1} \cdots \sum_{k_1=0}^{N_1-1} \exp(2\pi i k_L n_L / N_L) \times \cdots \\
\times \exp(2\pi i k_1 n_1 / N_1)\, h(k_1, \ldots, k_L)
\tag{12.4.3}
$$

where n_1 and k_1 range from 0 to $N_1 - 1$, $\ldots$, n_L and k_L range from 0 to $N_L - 1$. How many calls to a one-dimensional FFT are in (12.4.3)? Quite a few! For each value of $k_1, k_2, \ldots, k_{L-1}$ you FFT to transform the L index. Then for each value of $k_1, k_2, \ldots, k_{L-2}$ and n_L you FFT to transform the $L-1$ index. And so on. It is best to rely on someone else having done the bookkeeping for once and for all.

The inverse transforms of (12.4.1) or (12.4.3) are just what you would expect them to be: Change the i's in the exponentials to $-i$'s, and put an overall factor of $1/(N_1 \times \cdots \times N_L)$ in front of the whole thing. Most other features of multidimensional FFTs are also analogous to features already discussed in the one-dimensional case:

- Frequencies are arranged in wrap-around order in the transform, but now for each separate dimension.
- The input data are also treated as if they were wrapped around. If they are discontinuous across this periodic identification (in any dimension) then the spectrum will have some excess power at high frequencies because of the discontinuity. The fix, if you care, is to remove multidimensional linear trends.
- If you are doing spatial filtering and are worried about wrap-around effects, then you need to zero-pad all around the border of the multidimensional array. However, be sure to notice how costly zero-padding is in multidimensional transforms. If you use too thick a zero-pad, you are going to waste a *lot* of storage, especially in 3 or more dimensions!
- Aliasing occurs as always if sufficient bandwidth limiting does not exist along one or more of the dimensions of the transform.

$$f_2 = 0 \qquad f_2 = \frac{1}{N_2 \Delta_2} \qquad f_2 = \frac{N_2/2 - 1}{N_2 \Delta_2} \qquad f_2 = \pm \frac{1}{2\Delta_2} \qquad f_2 = -\frac{N_2/2 - 1}{N_2 \Delta_2} \qquad f_2 = -\frac{1}{N_2 \Delta_2}$$

$$\text{col. 1} \qquad \text{col. 2} \qquad \text{col. } \frac{N_2}{2} \qquad \text{col. } \frac{N_2}{2} + 1 \qquad \text{col. } \frac{N_2}{2} + 2 \qquad \text{col. } N_2$$
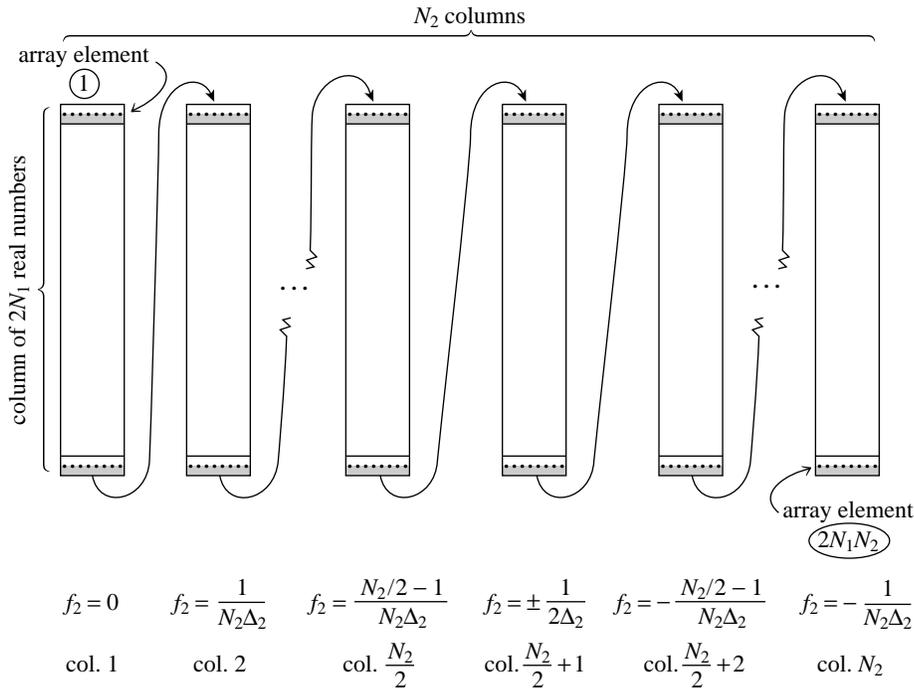
Figure 12.4.1. Storage arrangement of frequencies in the output $H(f_1, f_2)$ of a two-dimensional FFT. The input data is a two-dimensional $N_1 \times N_2$ array $h(t_1, t_2)$ (stored by columns of complex numbers). The output is also stored by complex columns. Each column corresponds to a particular value of f_2, as shown in the figure. Within each column, the arrangement of frequencies f_1 is exactly as shown in Figure 12.2.2 Δ_1 and Δ_2 are the sampling intervals in the 1 and 2 directions, respectively. The total number of (real) array elements is $2N_1N_2$. The program fourn can also do more than two dimensions, and the storage arrangement generalizes in the obvious way.

The routine fourn that we furnish herewith is a descendant of one written by N. M. Brenner. It requires as input (i) a scalar, telling the number of dimensions, e.g., 2; (ii) a vector, telling the length of the array in each dimension, e.g., (32,64). Note that these lengths *must all* be powers of 2, and are the numbers of *complex* values in each direction; (iii) the usual scalar equal to ± 1 indicating whether you want the transform or its inverse; and, finally (iv) the array of data.

A few words about the data array: fourn accesses it as a one-dimensional array of real numbers, of length equal to twice the product of the lengths of the L dimensions. It assumes that the array represents an L-dimensional complex array, in normal FORTRAN order. Normal FORTRAN order means: (i) each complex value occupies two sequential locations, real part followed by imaginary; (ii) the first subscript changes most rapidly as one goes through the array; the last subscript changes least rapidly; (iii) subscripts range from 1 to their maximum values $(N_1, N_2, \ldots, N_L,$ respectively), rather than from 0 to $N_1 - 1$, $N_2 - 1, \ldots, N_L - 1$. Almost all failures to get fourn to work result from improper understanding of the above ordering of the data array, so take care! (Figure 12.4.1 illustrates the format of the output array.)

```
      SUBROUTINE fourn(data,nn,ndim,isign)
      INTEGER isign,ndim,nn(ndim)
      REAL data(*)
```
    Replaces data by its ndim-dimensional discrete Fourier transform, if isign is input as
    1. nn(1:ndim) is an integer array containing the lengths of each dimension (number of
    complex values), which MUST all be powers of 2. data is a real array of length twice the
    product of these lengths, in which the data are stored as in a multidimensional complex
    FORTRAN array. If isign is input as −1, data is replaced by its inverse transform times
    the product of the lengths of all dimensions.
```
      INTEGER i1,i2,i2rev,i3,i3rev,ibit,idim,ifp1,ifp2,ip1,ip2,
     *     ip3,k1,k2,n,nprev,nrem,ntot
      REAL tempi,tempr
      DOUBLE PRECISION theta,wi,wpi,wpr,wr,wtemp        Double precision for trigonometric re-
      ntot=1                                                       currences.
      do 11 idim=1,ndim                 Compute total number of complex values.
          ntot=ntot*nn(idim)
      enddo 11
      nprev=1
      do 18 idim=1,ndim                 Main loop over the dimensions.
          n=nn(idim)
          nrem=ntot/(n*nprev)
          ip1=2*nprev
          ip2=ip1*n
          ip3=ip2*nrem
          i2rev=1
          do 14 i2=1,ip2,ip1            This is the bit-reversal section of the routine.
              if(i2.lt.i2rev)then
                  do 13 i1=i2,i2+ip1-2,2
                      do 12 i3=i1,ip3,ip2
                          i3rev=i2rev+i3-i2
                          tempr=data(i3)
                          tempi=data(i3+1)
                          data(i3)=data(i3rev)
                          data(i3+1)=data(i3rev+1)
                          data(i3rev)=tempr
                          data(i3rev+1)=tempi
                      enddo 12
                  enddo 13
              endif
              ibit=ip2/2
1             if ((ibit.ge.ip1).and.(i2rev.gt.ibit)) then
                  i2rev=i2rev-ibit
                  ibit=ibit/2
              goto 1
              endif
              i2rev=i2rev+ibit
          enddo 14
          ifp1=ip1                      Here begins the Danielson-Lanczos section of the routine.
2         if(ifp1.lt.ip2)then
              ifp2=2*ifp1
              theta=isign*6.28318530717959d0/(ifp2/ip1)     Initialize for the trig. recur-
              wpr=-2.d0*sin(0.5d0*theta)**2                          rence.
              wpi=sin(theta)
              wr=1.d0
              wi=0.d0
              do 17 i3=1,ifp1,ip1
                  do 16 i1=i3,i3+ip1-2,2
                      do 15 i2=i1,ip3,ifp2
                          k1=i2          Danielson-Lanczos formula:
                          k2=k1+ifp1
                          tempr=sngl(wr)*data(k2)-sngl(wi)*data(k2+1)
                          tempi=sngl(wr)*data(k2+1)+sngl(wi)*data(k2)
                          data(k2)=data(k1)-tempr
                          data(k2+1)=data(k1+1)-tempi
```

```
                data(k1)=data(k1)+tempr
                data(k1+1)=data(k1+1)+tempi
            enddo 15
        enddo 16
        wtemp=wr            Trigonometric recurrence.
        wr=wr*wpr-wi*wpi+wr
        wi=wi*wpr+wtemp*wpi+wi
    enddo 17
    ifp1=ifp2
    goto 2
    endif
    nprev=n*nprev
enddo 18
return
END
```

CITED REFERENCES AND FURTHER READING:

Nussbaumer, H.J. 1982, *Fast Fourier Transform and Convolution Algorithms* (New York: Springer-Verlag).

12.5 Fourier Transforms of Real Data in Two and Three Dimensions

Two-dimensional FFTs are particularly important in the field of image processing. An image is usually represented as a two-dimensional array of pixel intensities, real (and usually positive) numbers. One commonly desires to filter high, or low, frequency spatial components from an image; or to convolve or deconvolve the image with some instrumental point spread function. Use of the FFT is by far the most efficient technique.

In three dimensions, a common use of the FFT is to solve Poisson's equation for a potential (e.g., electromagnetic or gravitational) on a three-dimensional lattice that represents the discretization of three-dimensional space. Here the source terms (mass or charge distribution) and the desired potentials are also real. In two and three dimensions, with large arrays, memory is often at a premium. It is therefore important to perform the FFTs, insofar as possible, on the data "in place." We want a routine with functionality similar to the multidimensional FFT routine `fourn` (§12.4), but which operates on real, not complex, input data. We give such a routine in this section. The development is analogous to that of §12.3 leading to the one-dimensional routine `realft`. (You might wish to review that material at this point, particularly equation 12.3.5.)

It is convenient to think of the independent variables $n_1, \ldots, n_L$ in equation (12.4.3) as representing an L-dimensional vector $\vec{n}$ in wave-number space, with values on the lattice of integers. The transform $H(n_1, \ldots, n_L)$ is then denoted $H(\vec{n})$.

It is easy to see that the transform $H(\vec{n})$ is periodic in each of its L dimensions. Specifically, if $\vec{P}_1, \vec{P}_2, \vec{P}_3, \ldots$ denote the vectors $(N_1, 0, 0, \ldots)$, $(0, N_2, 0, \ldots)$, $(0, 0, N_3, \ldots)$, and so forth, then

$$H(\vec{n} \pm \vec{P}_j) = H(\vec{n}) \qquad j = 1, \ldots, L \tag{12.5.1}$$

Equation (12.5.1) holds for any input data, real or complex. When the data is real, we have the additional symmetry

$$H(-\vec{n}) = H(\vec{n})^* \qquad (12.5.2)$$

Equations (12.5.1) and (12.5.2) imply that the full transform can be trivially obtained from the subset of lattice values $\vec{n}$ that have

$$
\begin{aligned}
0 &\leq n_1 \leq \frac{N_1}{2} \\
0 &\leq n_2 \leq N_2 - 1 \\
&\cdots \\
0 &\leq n_L \leq N_L - 1
\end{aligned}
\qquad (12.5.3)
$$

In fact, this set of values is overcomplete, because there are additional symmetry relations among the transform values that have $n_1 = 0$ and $n_1 = N_1/2$. However these symmetries are complicated and their use becomes extremely confusing. Therefore, we will compute our FFT on the lattice subset of equation (12.5.3), even though this requires a small amount of extra storage for the answer, i.e., the transform is not *quite* "in place." (Although an in-place transform is in fact possible, we have found it virtually impossible to explain to any user how to unscramble its output, i.e., where to find the real and imaginary components of the transform at some particular frequency!)

Figure 12.5.1 shows the storage scheme that we will use for the input data and the output transform. The figure is specialized to the case of two dimensions, $L = 2$, but the generalization to higher dimensions is obvious. The input data is a two-dimensional real array of dimensions N_1 (called nn1) by N_2 (called nn2). Notice that the FORTRAN subscripts number from 1 to nn1, and not from 0 to $N_1 - 1$. The output spectrum is in two complex arrays, one two-dimensional and the other one-dimensional. The two-dimensional one, spec, has dimensions nn1/2 by nn2. This is exactly half the size of the input data array; but since it is complex, it is the same amount of storage. In fact, spec will share storage with (and overwrite) the input data array. As the figure shows, spec contains those spectral components whose first component of frequency, f_1, ranges from zero to just short of the Nyquist frequency f_c. The full range of positive and negative second-component of frequencies, f_2, is stored, in wrap-around order (see §12.2), with negative frequencies shifted by exactly one period to put them "above" the positive frequencies, as the figure indicates. The figure also indicates how the additional $L - 1$ (here, one-) dimensional array speq stores only that single value of n_1 that corresponds to the Nyquist frequency, but all values of n_2, etc.

With this much introduction, the implementing procedure, called rlft3, is something of an anticlimax. The routine is written for the case of $L = 3$ dimensions, but (we will explain below) it can be used without modification for $L = 2$ also; and it is quite trivial to generalize it to larger L. Look at the innermost ("do $_{13}$") loop in the procedure, and you will see equation (12.3.5) implemented on the *first* transform index. The case of i1=1 is coded separately, to account for the fact that speq is to be filled instead of spec (which is here called data since it shares storage with

Figure 12.5.1. Input and output data arrangement for `rlft3` in the case of two-dimensional data. The input data array is a real, two-dimensional array. The output data array `spec` is a complex, two-dimensional array whose $(1, 1)$ element contains the $f_1 = f_2 = 0$ spectral component; a complete set of f_2 values are stored in wrap-around order, while only positive f_1 values are stored (others being obtainable by symmetry). The output array `speq` contains components with f_1 equal to the Nyquist frequency.

the input array). The three enclosing do loops (indices `i2`, `i1`, and `i3`, from inside to outside) could in fact be done in any order — their actions all commute. We chose the order shown because of the following considerations: (i) `i1` should not be the inner loop, because if it is, then the recurrence relations on `wr` and `wi` become burdensome. (ii) On virtual-memory machines, `i3` should be the outer loop, because (with FORTRAN order of array storage) this results in the array `data`, which might be very large, being accessed in block sequential order.

Note that the work done in `rlft3` is quite (logarithmically) small, compared to the associated complex FFT, `fourn`. For this reason, we allow ourselves the clarity of using FORTRAN complex arithmetic even when (as in the multiplications by `c1` and `c2`) there are a few unnecessary operations. The routine `rlft3` is based on an earlier routine by G.B. Rybicki.

```
      SUBROUTINE rlft3(data,speq,nn1,nn2,nn3,isign)
      INTEGER isign,nn1,nn2,nn3
      COMPLEX data(nn1/2,nn2,nn3),speq(nn2,nn3)
C     USES fourn
          Given a two- or three-dimensional real array data whose dimensions are nn1, nn2, nn3
          (where nn3 is 1 for the case of a two-dimensional array), this routine returns (for isign=1)
          the complex fast Fourier transform as two complex arrays: On output, data contains the
          zero and positive frequency values of the first frequency component, while speq contains
          the Nyquist critical frequency values of the first frequency component. Second (and third)
          frequency components are stored for zero, positive, and negative frequencies, in standard
          wrap-around order. For isign=-1, the inverse transform (times nn1*nn2*nn3/2 as a
          constant multiplicative factor) is performed, with output data (viewed as a real array)
          deriving from input data (viewed as complex) and speq. For inverse transforms on data
          not generated first by a forward transform, make sure the complex input data array satisfies
          property (12.5.2). The dimensions nn1, nn2, nn3 must always be integer powers of 2.
      INTEGER i1,i2,i3,j1,j2,j3,nn(3)
      DOUBLE PRECISION theta,wi,wpi,wpr,wr,wtemp
      COMPLEX c1,c2,h1,h2,w              Note that data is dimensioned as complex, its output
      c1=cmplx(0.5,0.0)                       format.
      c2=cmplx(0.0,-0.5*isign)
      theta=6.28318530717959d0/dble(isign*nn1)
      wpr=-2.0d0*sin(0.5d0*theta)**2
      wpi=sin(theta)
      nn(1)=nn1/2
      nn(2)=nn2
      nn(3)=nn3
      if(isign.eq.1)then                Case of forward transform.
          call fourn(data,nn,3,isign) Here is where most all of the compute time is spent.
          do 12 i3=1,nn3               Extend data periodically into speq.
              do 11 i2=1,nn2
                  speq(i2,i3)=data(1,i2,i3)
              enddo 11
          enddo 12
      endif
      do 15 i3=1,nn3
          j3=1
          if (i3.ne.1) j3=nn3-i3+2       Zero frequency is its own reflection, otherwise locate cor-
          wr=1.0d0                          responding negative frequency in wrap-around order.
          wi=0.0d0                       Initialize trigonometric recurrence.
          do 14 i1=1,nn1/4+1
              j1=nn1/2-i1+2
              do 13 i2=1,nn2
                  j2=1
                  if (i2.ne.1) j2=nn2-i2+2
                  if(i1.eq.1)then       Equation (12.3.5).
                      h1=c1*(data(1,i2,i3)+conjg(speq(j2,j3)))
                      h2=c2*(data(1,i2,i3)-conjg(speq(j2,j3)))
```

Figure 12.5.2.   (a) A two-dimensional image with intensities either purely black or purely white. (b) The same image, after it has been low-pass filtered using `rlft3`. Regions with fine-scale features become gray.

```
                data(1,i2,i3)=h1+h2
                speq(j2,j3)=conjg(h1-h2)
            else
                h1=c1*(data(i1,i2,i3)+conjg(data(j1,j2,j3)))
                h2=c2*(data(i1,i2,i3)-conjg(data(j1,j2,j3)))
                data(i1,i2,i3)=h1+w*h2
                data(j1,j2,j3)=conjg(h1-w*h2)
            endif
        enddo 13
        wtemp=wr                    Do the recurrence.
        wr=wr*wpr-wi*wpi+wr
        wi=wi*wpr+wtemp*wpi+wi
        w=cmplx(sngl(wr),sngl(wi))
    enddo 14
enddo 15
if(isign.eq.-1)then             Case of reverse transform.
    call fourn(data,nn,3,isign)
endif
return
END
```

We now give some fragments from notional calling programs, to clarify the use of `rlft3` for two- and three-dimensional data. Note that the routine does not actually distinguish between two and three dimensions; two is treated like three, but with the third dimension having length 1. Since the third dimension is the outer loop, almost no inefficiency is introduced.

The first program fragment FFTs a two-dimensional data array, allows for some processing on it, e.g., filtering, and then takes the inverse transform. Figure 12.5.2 shows an example of the use of this kind of code: A sharp image becomes blurry when its high-frequency spatial components are suppressed by the factor (here) $\max{(1 - 6f^2/f_c^2, 0)}$. The second program example illustrates a three-dimensional transform, where the three dimensions have different lengths. The third program example is an example of convolution, as it might occur in a program to compute the potential generated by a three-dimensional distribution of sources.

```
      PROGRAM exmpl1
         This fragment shows how one might filter a 256 by 256 digital image.
      INTEGER N1,N2,N3
      PARAMETER (N1=256,N2=256,N3=1)  Note that the third component must be set to 1.
C     USES rlft3
      REAL data(N1,N2)
      COMPLEX spec(N1/2,N2),speq(N2)
      EQUIVALENCE (data,spec)
C     ...                            Here the image would be loaded into data.
      call rlft3(data,speq,N1,N2,N3,1)
C     ...                            Here the arrays spec and speq would be multiplied by a suit-
      call rlft3(data,speq,N1,N2,N3,-1)      able filter function (of frequency).
C     ...                            Here the filtered image would be unloaded from data.
      END


      PROGRAM exmpl2
         This fragment shows how one might FFT a real three-dimensional array of size 32 by 64
         by 16.
      INTEGER N1,N2,N3
      PARAMETER (N1=32,N2=64,N3=16)
C     USES rlft3
      REAL data(N1,N2,N3)
      COMPLEX spec(N1/2,N2,N3),speq(N2,N3)
      EQUIVALENCE (data,spec)
C     ...                            Here load data.
      call rlft3(data,speq,N1,N2,N3,1)
C     ...                            Here unload spec and speq.
      END


      PROGRAM exmpl3
         This fragment shows how one might convolve two real, three-dimensional arrays of size 32
         by 32 by 32, replacing the first array by the result.
      INTEGER N
      PARAMETER (N=32)
C     USES rlft3
      INTEGER j
      REAL fac,data1(N,N,N),data2(N,N,N)
      COMPLEX spec1(N/2,N,N),speq1(N,N),spec2(N/2,N,N),speq2(N,N),
     *    zpec1(N*N*N/2),zpeq1(N*N),zpec2(N*N*N/2),zpeq2(N*N)
      EQUIVALENCE (data1,spec1,zpec1), (data2,spec2,zpec2),
     *    (speq1,zpeq1), (speq2,zpeq2)
C     ...
      call rlft3(data1,speq1,N,N,N,1)          FFT both input arrays.
      call rlft3(data2,speq2,N,N,N,1)
      fac=2./(N*N*N)                           Factor needed to get normalized inverse.
      do 11 j=1,N*N*N/2                         The sole purpose of the zpecs and zpeqs is to make
         zpec1(j)=fac*zpec1(j)*zpec2(j)         this a single do-loop instead of three-nested ones.
      enddo 11
      do 12 j=1,N*N
         zpeq1(j)=fac*zpeq1(j)*zpeq2(j)
      enddo 12
      call rlft3(data1,speq1,N,N,N,-1)   Inverse FFT the product of the two FFTs.
C     ...
      END
```

To extend `rlft3` to four dimensions, you simply add an additional (outer) nested
do loop in `i4`, analogous to the present `i3`. (Modifying the routine to do an *arbitrary*
number of dimensions, as in `fourn`, is a good programming exercise for the reader.)

CITED REFERENCES AND FURTHER READING:

Brigham, E.O. 1974, *The Fast Fourier Transform* (Englewood Cliffs, NJ: Prentice-Hall).

Swartztrauber, P. N. 1986, *Mathematics of Computation*, vol. 47, pp. 323–346.

12.6 External Storage or Memory-Local FFTs

Sometime in your life, you might have to compute the Fourier transform of a *really large* data set, larger than the size of your computer's physical memory. In such a case, the data will be stored on some external medium, such as magnetic or optical tape or disk. Needed is an algorithm that makes some manageable number of sequential passes through the external data, processing it on the fly and outputting intermediate results to other external media, which can be read on subsequent passes.

In fact, an algorithm of just this description was developed by Singleton [1] very soon after the discovery of the FFT. The algorithm requires four sequential storage devices, each capable of holding half of the input data. The first half of the input data is initially on one device, the second half on another.

Singleton's algorithm is based on the observation that it is possible to bit-reverse 2^M values by the following sequence of operations: On the first pass, values are read alternately from the two input devices, and written to a single output device (until it holds half the data), and then to the other output device. On the second pass, the output devices become input devices, and vice versa. Now, we copy *two* values from the first device, then *two* values from the second, writing them (as before) first to fill one output device, then to fill a second. Subsequent passes read 4, 8, etc., input values at a time. After completion of pass $M - 1$, the data are in bit-reverse order.

Singleton's next observation is that it is possible to alternate the passes of essentially this bit-reversal technique with passes that implement one stage of the Danielson-Lanczos combination formula (12.2.3). The scheme, roughly, is this: One starts as before with half the input data on one device, half on another. In the first pass, one complex value is read from each input device. Two combinations are formed, and one is written to each of two output devices. After this "computing" pass, the devices are rewound, and a "permutation" pass is performed, where groups of values are read from the first input device and alternately written to the first and second output devices; when the first input device is exhausted, the second is similarly processed. This sequence of computing and permutation passes is repeated $M - K - 1$ times, where 2^K is the size of internal buffer available to the program. The second phase of the computation consists of a final K computation passes. What distinguishes the second phase from the first is that, now, the permutations are local enough to do in place during the computation. There are thus no separate permutation passes in the second phase. In all, there are $2M - K - 2$ passes through the data.

Here is an implementation of Singleton's algorithm, based on [1]:

```
      SUBROUTINE fourfs(iunit,nn,ndim,isign)
      INTEGER ndim,nn(ndim),isign,iunit(4),KBF
      PARAMETER (KBF=128)
C     USES fourew
            One- or multi-dimensional Fourier transform of a large data set stored on external media.
            On input, ndim is the number of dimensions, and nn(1:ndim) contains the lengths of
            each dimension (number of complex values), which must be powers of two. iunit(1:4)
            contains the unit numbers of 4 sequential files, each large enough to hold half of the data.
            The four units must be opened for FORTRAN unformatted access. The input data must be
            in FORTRAN normal order, with its first half stored on unit iunit(1), its second half on
            iunit(2), in unformatted form, with KBF real numbers per record. isign should be set
            to 1 for the Fourier transform, to −1 for its inverse. On output, values in the array iunit
            may have been permuted; the first half of the result is stored on iunit(3), the second
            half on iunit(4). N.B.: For ndim > 1, the output is stored by rows, i.e., *not* in FORTRAN
            normal order; in other words, the output is the transpose of that which would have been
            produced by routine fourn.
      INTEGER j,j12,jk,k,kk,n,mm,kc,kd,ks,kr,nr,ns,nv,jx,
     *      mate(4),na,nb,nc,nd
      REAL tempr,tempi,afa(KBF),afb(KBF),afc(KBF)
      DOUBLE PRECISION wr,wi,wpr,wpi,wtemp,theta
      SAVE mate
      DATA mate /2,1,4,3/
```

```
      n=1
      do 11 j=1,ndim
          n=n*nn(j)
          if (nn(j).le.1)
*             pause 'invalid dimension or wrong ndim in fourfs'
      enddo 11
      nv=ndim
      jk=nn(nv)
      mm=n
      ns=n/KBF
      nr=ns/2
      kc=0
      kd=KBF/2
      ks=n
      call fourew(iunit,na,nb,nc,nd)
      The first phase of the transform starts here.
1     continue                          Start of the computing pass.
          theta=3.141592653589793d0/(isign*n/mm)
          wpr=-2.d0*sin(0.5d0*theta)**2
          wpi=sin(theta)
          wr=1.d0
          wi=0.d0
          mm=mm/2
          do 13 j12=1,2
              kr=0
2             continue
                  read (iunit(na)) (afa(jx),jx=1,KBF)
                  read (iunit(nb)) (afb(jx),jx=1,KBF)
                  do 12 j=1,KBF,2
                      tempr=sngl(wr)*afb(j)-sngl(wi)*afb(j+1)
                      tempi=sngl(wi)*afb(j)+sngl(wr)*afb(j+1)
                      afb(j)=afa(j)-tempr
                      afa(j)=afa(j)+tempr
                      afb(j+1)=afa(j+1)-tempi
                      afa(j+1)=afa(j+1)+tempi
                  enddo 12
                  kc=kc+kd
                  if (kc.eq.mm) then
                      kc=0
                      wtemp=wr
                      wr=wr*wpr-wi*wpi+wr
                      wi=wi*wpr+wtemp*wpi+wi
                  endif
                  write (iunit(nc)) (afa(jx),jx=1,KBF)
                  write (iunit(nd)) (afb(jx),jx=1,KBF)
              kr=kr+1
              if (kr.lt.nr) goto 2
              if(j12.eq.1.and.ks.ne.n.and.ks.eq.KBF) then
                  na=mate(na)
                  nb=na
              endif
              if (nr.eq.0) goto 3
          enddo 13
3         call fourew(iunit,na,nb,nc,nd)          Start of the permutation pass.
          jk=jk/2
4         if (jk.eq.1) then
              mm=n
              nv=nv-1
              jk=nn(nv)
          goto 4
          endif
          ks=ks/2
          if (ks.gt.KBF) then
              do 16 j12=1,2
```

```fortran
            do 15 kr=1,ns,ks/KBF
                do 14 k=1,ks,KBF
                    read (iunit(na)) (afa(jx),jx=1,KBF)
                    write (iunit(nc)) (afa(jx),jx=1,KBF)
                enddo 14
                nc=mate(nc)
            enddo 15
            na=mate(na)
        enddo 16
        call fourew(iunit,na,nb,nc,nd)
        goto 1
      else if (ks.eq.KBF) then
            nb=na
            goto 1
      endif
    continue
    j=1
```

The second phase of the transform starts here. Now, the remaining permutations are sufficiently local to be done in place.

```fortran
5   continue
        theta=3.141592653589793d0/(isign*n/mm)
        wpr=-2.d0*sin(0.5d0*theta)**2
        wpi=sin(theta)
        wr=1.d0
        wi=0.d0
        mm=mm/2
        ks=kd
        kd=kd/2
        do 18 j12=1,2
            do 17 kr=1,ns
                read (iunit(na)) (afc(jx),jx=1,KBF)
                kk=1
                k=ks+1
6           continue
                tempr=sngl(wr)*afc(kk+ks)-sngl(wi)*afc(kk+ks+1)
                tempi=sngl(wi)*afc(kk+ks)+sngl(wr)*afc(kk+ks+1)
                afa(j)=afc(kk)+tempr
                afb(j)=afc(kk)-tempr
                afa(j+1)=afc(kk+1)+tempi
                afb(j+1)=afc(kk+1)-tempi
                j=j+2
                kk=kk+2
            if (kk.lt.k) goto 6
            kc=kc+kd
            if (kc.eq.mm) then
                kc=0
                wtemp=wr
                wr=wr*wpr-wi*wpi+wr
                wi=wi*wpr+wtemp*wpi+wi
            endif
            kk=kk+ks
            if (kk.le.KBF) then
                k=kk+ks
                goto 6
            endif
            if (j.gt.KBF) then
                write (iunit(nc)) (afa(jx),jx=1,KBF)
                write (iunit(nd)) (afb(jx),jx=1,KBF)
                j=1
            endif
            enddo 17
            na=mate(na)
        enddo 18
        call fourew(iunit,na,nb,nc,nd)
```

```
      jk=jk/2
  if (jk.gt.1) goto 5
  mm=n
7 if (nv.gt.1) then
      nv=nv-1
      jk=nn(nv)
      if (jk.eq.1) goto 7
      goto 5
  endif
  return
  END
```

```
  SUBROUTINE fourew(iunit,na,nb,nc,nd)
  INTEGER na,nb,nc,nd,iunit(4),ii
      Utility used by fourfs. Rewinds and renumbers the four files.
  do 11 ii=1,4
      rewind(unit=iunit(ii))
  enddo 11
  ii=iunit(2)
  iunit(2)=iunit(4)
  iunit(4)=ii
  ii=iunit(1)
  iunit(1)=iunit(3)
  iunit(3)=ii
  na=3
  nb=4
  nc=1
  nd=2
  return
  END
```

For one-dimensional data, Singleton's algorithm produces output in exactly the same order as a standard FFT (e.g., four1). For multidimensional data, the output is the *transpose* of the conventional arrangement (e.g., the output of fourn). This peculiarity, which is intrinsic to the method, is generally only a minor inconvenience. For convolutions, one simply computes the component-by-component product of two transforms in their nonstandard arrangement, and then does an inverse transform on the result. Note that, if the lengths of the different dimensions are not all the same, then you must reverse the order of the values in nn(1:ndim) (thus giving the transpose dimensions) before performing the inverse transform. Note also that, just like fourn, performing a transform and then an inverse results in multiplying the original data by the product of the lengths of all dimensions.

We leave it as an exercise for the reader to figure out how to reorder fourfs's output into normal order, taking additional passes through the externally stored data. We doubt that such reordering is ever really needed.

You will likely want to modify fourfs to fit your particular application. For example, as written, KBF $\equiv 2^K$ plays the dual role of being the size of the internal buffers, and the record size of the unformatted reads and writes. The latter role limits its size to that allowed by your machine's I/O facility. It is a simple matter to perform multiple reads for a much larger KBF, thus reducing the number of passes by a few.

Another modification of fourfs would be for the case where your virtual memory machine has sufficient address space, but not sufficient physical memory, to do an efficient FFT by the conventional algorithm (whose memory references are extremely nonlocal). In that case, you will need to replace the reads, writes, and rewinds by mappings of the arrays afa, afb, and afc into your address space. In other words, these arrays are replaced by references to a single data array, with offsets that get modified wherever fourfs performs an I/O operation. The resulting algorithm will have its memory references local within blocks of size KBF. Execution speed is thereby sometimes increased enormously, albeit at the cost of requiring twice as much virtual memory as an in-place FFT.

CITED REFERENCES AND FURTHER READING:

Singleton, R.C. 1967, *IEEE Transactions on Audio and Electroacoustics*, vol. AU-15, pp. 91–97. [1]

Oppenheim, A.V., and Schafer, R.W. 1989, *Discrete-Time Signal Processing* (Englewood Cliffs, NJ: Prentice-Hall), Chapter 9.