1. GENERAL RELATIONSHIPS AND TECHNIQUES

Whenever the functions to which the Fourier transform is applied are band-limited, or can be well approximated by band-limited functions, the discrete Fourier transform (DFT) provides a means of constructing explicit numerical solutions to the problems at hand. A great variety of investigations in physics, engineering and applied mathematics thus lead to DFT calculations, to such a degree that, at the time of writing, about 50% of all supercomputer CPU time is alleged to be spent calculating DFTs.

The straightforward use of the defining formulae for the DFT leads to calculations of size N^2 for N sample points, which become unfeasible for any but the smallest problems. Much ingenuity has therefore been exerted on the design and implementation of faster algorithms for calculating the DFT (McClellan & Rader, 1979; Nussbaumer, 1981; Blahut, 1985; Brigham, 1988). The most famous is that of Cooley & Tukey (1965) which heralded the age of digital signal processing. However, it had been preceded by the prime factor algorithm of Good (1958, 1960), which has lately been the basis of many new developments. Recent historical research (Goldstine, 1977, pp. 249–253; Heideman *et al.*, 1984) has shown that Gauss essentially knew the Cooley–Tukey algorithm as early as 1805 (before Fourier's 1807 work on harmonic analysis!); while it has long been clear that Dirichlet knew of the basis of the prime factor algorithm and used it extensively in his theory of multiplicative characters [see *e.g.* Chapter I of Ayoub (1963), and Chapters 6 and 8 of Apostol (1976)]. Thus the computation of the DFT, far from being a purely technical and rather narrow piece of specialized numerical analysis, turns out to have very rich connections with such central areas of pure mathematics as number theory (algebraic and analytic), the representation theory of certain Lie groups and coding theory – to list only a few. The interested reader may consult Auslander & Tolimieri (1979); Auslander, Feig & Winograd (1982, 1984); Auslander & Tolimieri (1985); Tolimieri (1985).

One-dimensional algorithms are examined first. The Sande mixed-radix version of the Cooley–Tukey algorithm only calls upon the additive structure of congruence classes of integers. The prime factor algorithm of Good begins to exploit some of their multiplicative structure, and the use of relatively prime factors leads to a stronger factorization than that of Sande. Fuller use of the multiplicative structure, *via* the group of units, leads to the Rader algorithm; and the factorization of short convolutions then yields the Winograd algorithms.

Multidimensional algorithms are at first built as tensor products of one-dimensional elements. The problem of factoring the DFT in several dimensions simultaneously is then examined. The section ends with a survey of attempts at formalizing the interplay between algorithm structure and computer architecture for the purpose of automating the design of optimal DFT code.

It was originally intended to incorporate into this section a survey of all the basic notions and results of abstract algebra which are called upon in the course of these developments, but time limitations have made this impossible. This material, however, is adequately covered by the first chapter of Tolimieri *et al.* (1989) in a form tailored for the same purposes. Similarly, the inclusion of numerous detailed examples of the algorithms described here has had to be postponed to a later edition, but an abundant supply of such examples may be found in the signal processing literature, for instance in the books by McClellan & Rader (1979), Blahut (1985), and Tolimieri *et al.* (1989).

1.3.3.2. *One-dimensional algorithms*

Throughout this section we will denote by $e(t)$ the expression $\exp(2\pi i t)$, $t \in \mathbb{R}$. The mapping $t \longmapsto e(t)$ has the following properties:

$$\begin{aligned} e(t_1 + t_2) &= e(t_1)e(t_2) \\ e(-t) &= \overline{e(t)} = [e(t)]^{-1} \\ e(t) &= 1 \Leftrightarrow t \in \mathbb{Z}. \end{aligned}$$

Thus e defines an isomorphism between the additive group $\mathbb{R}/\mathbb{Z}$ (the reals modulo the integers) and the multiplicative group of complex numbers of modulus 1. It follows that the mapping $\ell \longmapsto e(\ell/N)$, where $\ell \in \mathbb{Z}$ and N is a positive integer, defines an isomorphism between the one-dimensional residual lattice $\mathbb{Z}/N\mathbb{Z}$ and the multiplicative group of Nth roots of unity.

The DFT on N points then relates vectors $\mathbf{X}$ and $\mathbf{X}^*$ in W and W^* through the linear transformations:

$$\begin{aligned} F(N): \quad & X(k) = \frac{1}{N} \sum_{k^* \in \mathbb{Z}/N\mathbb{Z}} X^*(k^*) e(-k^*k/N) \\ \bar{F}(N): \quad & X^*(k^*) = \sum_{k \in \mathbb{Z}/N\mathbb{Z}} X(k) e(k^*k/N). \end{aligned}$$

1.3.3.2.1. *The Cooley–Tukey algorithm*

The presentation of Gentleman & Sande (1966) will be followed first [see also Cochran *et al.* (1967)]. It will then be reinterpreted in geometric terms which will prepare the way for the treatment of multidimensional transforms in Section 1.3.3.3.

Suppose that the number of sample points N is composite, say $N = N_1N_2$. We may write k to the base N_1 and k^* to the base N_2 as follows:

$$\begin{aligned} k &= k_1 + N_1k_2 \quad k_1 \in \mathbb{Z}/N_1\mathbb{Z}, \quad k_2 \in \mathbb{Z}/N_2\mathbb{Z} \\ k^* &= k_2^* + k_1^*N_2 \quad k_1^* \in \mathbb{Z}/N_1\mathbb{Z}, \quad k_2^* \in \mathbb{Z}/N_2\mathbb{Z}. \end{aligned}$$

The defining relation for $\bar{F}(N)$ may then be written:

$$\begin{aligned} X^*(k_2^* + k_1^*N_2) = \sum_{k_1 \in \mathbb{Z}/N_1\mathbb{Z}} \sum_{k_2 \in \mathbb{Z}/N_2\mathbb{Z}} & X(k_1 + N_1k_2) \\ & \times e\left[\frac{(k_2^* + k_1^*N_2)(k_1 + N_1k_2)}{N_1N_2}\right]. \end{aligned}$$

The argument of $e[.]$ may be expanded as

$$\frac{k_2^*k_1}{N} + \frac{k_1^*k_1}{N_1} + \frac{k_2^*k_2}{N_2} + k_1^*k_2,$$

and the last summand, being an integer, may be dropped:

$$\begin{aligned} & X^*(k_2^* + k_1^*N_2) \\ & \quad = \sum_{k_1} \left\{ e\left(\frac{k_2^*k_1}{N}\right) \left[\sum_{k_2} X(k_1 + N_1k_2) e\left(\frac{k_2^*k_2}{N_2}\right) \right] \right\} \\ & \qquad \times e\left(\frac{k_1^*k_1}{N_1}\right). \end{aligned}$$

This computation may be decomposed into five stages, as follows:

(i) form the N_1 vectors $\mathbf{Y}_{k_1}$ of length N_2 by the prescription

$$Y_{k_1}(k_2) = X(k_1 + N_1k_2), \quad k_1 \in \mathbb{Z}/N_1\mathbb{Z}, \quad k_2 \in \mathbb{Z}/N_2\mathbb{Z};$$

(ii) calculate the N_1 transforms $\mathbf{Y}_{k_1}^*$ on N_2 points:

$$\mathbf{Y}_{k_1}^* = \bar{F}(N_2)[\mathbf{Y}_{k_1}], \quad k_1 \in \mathbb{Z}/N_1\mathbb{Z};$$

(iii) form the N_2 vectors $\mathbf{Z}_{k_2^*}$ of length N_1 by the prescription

$$\mathbf{Z}_{k_2^*}(k_1) = e\left(\frac{k_2^*k_1}{N}\right) Y_{k_1}^*(k_2^*), \quad k_1 \in \mathbb{Z}/N_1\mathbb{Z}, \quad k_2^* \in \mathbb{Z}/N_2\mathbb{Z};$$

(iv) calculate the N_2 transforms $\mathbf{Z}^*_{k_2^*}$ on N_1 points:

$$\mathbf{Z}^*_{k_2^*} = \bar{F}(N_1)[\mathbf{Z}_{k_2^*}], \quad k_2^* \in \mathbb{Z}/N_2\mathbb{Z};$$

(v) collect $X^*(k_2^* + k_1^* N_2)$ as $Z^*_{k_2^*}(k_1^*)$.

If the intermediate transforms in stages (ii) and (iv) are performed *in place*, *i.e.* with the results overwriting the data, then at stage (v) the result $X^*(k_2^* + k_1^* N_2)$ will be found at address $k_1^* + N_1 k_2^*$. This phenomenon is called *scrambling* by 'digit reversal', and stage (v) is accordingly known as *unscrambling*.

The initial *N-point* transform $\bar{F}(N)$ has thus been performed as N_1 transforms $\bar{F}(N_2)$ on N_2 points, followed by N_2 transforms $\bar{F}(N_1)$ on N_1 points, thereby reducing the arithmetic cost from $(N_1 N_2)^2$ to $N_1 N_2 (N_1 + N_2)$. The phase shifts applied at stage (iii) are traditionally called 'twiddle factors', and the transposition between k_1 and k_2^* can be performed by the fast recursive technique of Eklundh (1972). Clearly, this procedure can be applied recursively if N_1 and N_2 are themselves composite, leading to an overall arithmetic cost of order $N \log N$ if N has no large prime factors.

The Cooley–Tukey factorization may also be derived from a geometric rather than arithmetic argument. The decomposition $k = k_1 + N_1 k_2$ is associated to a geometric partition of the residual lattice $\mathbb{Z}/N\mathbb{Z}$ into N_1 copies of $\mathbb{Z}/N_2\mathbb{Z}$, each translated by $k_1 \in \mathbb{Z}/N_1\mathbb{Z}$ and 'blown up' by a factor N_1. This partition in turn induces a (direct sum) decomposition of $\mathbf{X}$ as

$$\mathbf{X} = \sum_{k_1} \mathbf{X}_{k_1},$$

where

$$\begin{aligned} X_{k_1}(k) &= X(k) \quad \text{if } k \equiv k_1 \bmod N_1, \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

According to (i), $\mathbf{X}_{k_1}$ is related to $\mathbf{Y}_{k_1}$ by *decimation by* N_1 and *offset by* k_1. By Section 1.3.2.7.2, $\bar{F}(N)[\mathbf{X}_{k_1}]$ is related to $\bar{F}(N_2)[\mathbf{Y}_{k_1}]$ by *periodization by* N_2 and *phase shift by* $e(k^* k_1/N)$, so that

$$X^*(k^*) = \sum_{k_1} e\left(\frac{k^* k_1}{N}\right) Y^*_{k_1}(k_2^*),$$

the periodization by N_2 being reflected by the fact that $Y^*_{k_1}$ does not depend on k_1^*. Writing $k^* = k_2^* + k_1^* N_2$ and expanding $k^* k_1$ shows that the phase shift contains both the twiddle factor $e(k_2^* k_1/N)$ and the kernel $e(k_1^* k_1/N_1)$ of $\bar{F}(N_1)$. The Cooley–Tukey algorithm is thus naturally associated to the coset decomposition of a lattice modulo a sublattice (Section 1.3.2.7.2).

It is readily seen that essentially the same factorization can be obtained for $F(N)$, up to the complex conjugation of the twiddle factors. The normalizing constant $1/N$ arises from the normalizing constants $1/N_1$ and $1/N_2$ in $F(N_1)$ and $F(N_2)$, respectively.

Factors of 2 are particularly simple to deal with and give rise to a characteristic computational structure called a 'butterfly loop'. If $N = 2M$, then two options exist:

(*a*) using $N_1 = 2$ and $N_2 = M$ leads to collecting the even-numbered coordinates of $\mathbf{X}$ into $\mathbf{Y}_0$ and the odd-numbered coordinates into $\mathbf{Y}_1$

$$\begin{aligned} Y_0(k_2) &= X(2k_2), \qquad k_2 = 0, \ldots, M-1, \\ Y_1(k_2) &= X(2k_2+1), \qquad k_2 = 0, \ldots, M-1, \end{aligned}$$

and writing:

$$\begin{aligned} X^*(k_2^*) &= Y_0^*(k_2^*) + e(k_2^*/N) Y_1^*(k_2^*), \\ &\qquad k_2^* = 0, \ldots, M-1; \\ X^*(k_2^* + M) &= Y_0^*(k_2^*) - e(k_2^*/N) Y_1^*(k_2^*), \\ &\qquad k_2^* = 0, \ldots, M-1. \end{aligned}$$

This is the original version of Cooley & Tukey, and the process of formation of $\mathbf{Y}_0$ and $\mathbf{Y}_1$ is referred to as 'decimation in time' (*i.e.* decimation along the *data* index $\mathbf{k}$).

(*b*) using $N_1 = M$ and $N_2 = 2$ leads to forming

$$\begin{aligned} Z_0(k_1) &= X(k_1) + X(k_1 + M), \qquad k_1 = 0, \ldots, M-1, \\ Z_1(k_1) &= [X(k_1) - X(k_1 + M)] e\left(\frac{k_1}{N}\right), \qquad k_1 = 0, \ldots, M-1, \end{aligned}$$

then obtaining separately the even-numbered and odd-numbered components of $\mathbf{X}^*$ by transforming $\mathbf{Z}_0$ and $\mathbf{Z}_1$:

$$\begin{aligned} X^*(2k_1^*) &= Z_0^*(k_1^*), \quad k_1^* = 0, \ldots, M-1; \\ X^*(2k_1^* + 1) &= Z_1^*(k_1^*), \quad k_1^* = 0, \ldots, M-1. \end{aligned}$$

This version is due to Sande (Gentleman & Sande, 1966), and the process of separately obtaining even-numbered and odd-numbered results has led to its being referred to as 'decimation in frequency' (*i.e.* decimation along the *result* index k^*).

By repeated factoring of the number N of sample points, the calculation of $F(N)$ and $\bar{F}(N)$ can be reduced to a succession of stages, the smallest of which operate on single prime factors of N. The reader is referred to Gentleman & Sande (1966) for a particularly lucid analysis of the programming considerations which help implement this factorization efficiently; see also Singleton (1969). Powers of two are often grouped together into factors of 4 or 8, which are advantageous in that they require fewer complex multiplications than the repeated use of factors of 2. In this approach, large prime factors P are detrimental, since they require a full P^2-size computation according to the defining formula.

1.3.3.2.2. *The Good (or prime factor) algorithm*

1.3.3.2.2.1. *Ring structure on* $\mathbb{Z}/N\mathbb{Z}$

The set $\mathbb{Z}/N\mathbb{Z}$ of congruence classes of integers modulo an integer N [see *e.g.* Apostol (1976), Chapter 5] inherits from $\mathbb{Z}$ not only the additive structure used in deriving the Cooley–Tukey factorization, but also a *multiplicative* structure in which the product of two congruence classes mod N is uniquely defined as the class of the ordinary product (in $\mathbb{Z}$) of representatives of each class. The multiplication can be distributed over addition in the usual way, endowing $\mathbb{Z}/N\mathbb{Z}$ with the structure of a *commutative ring*.

If N is composite, the ring $\mathbb{Z}/N\mathbb{Z}$ has *zero divisors*. For example, let $N = N_1 N_2$, let $n_1 \equiv N_1 \bmod N$, and let $n_2 \equiv N_2 \bmod N$: then $n_1 n_2 \equiv 0 \bmod N$. In the general case, a product of non-zero elements will be zero whenever these elements collect together all the factors of N. These circumstances give rise to a fundamental theorem in the theory of commutative rings, the *Chinese Remainder Theorem* (CRT), which will now be stated and proved [see Apostol (1976), Chapter 5; Schroeder (1986), Chapter 16].

1.3.3.2.2.2. *The Chinese remainder theorem*

Let $N = N_1 N_2 \ldots N_d$ be factored into a product of pairwise coprime integers, so that g.c.d. $(N_i, N_j) = 1$ for $i \neq j$. Then the system of congruence equations

$$\ell \equiv \ell_j \bmod N_j, \qquad j = 1, \ldots, d,$$

has a unique solution $\ell \bmod N$. In other words, each $\ell \in \mathbb{Z}/N\mathbb{Z}$ is

associated in a one-to-one fashion to the d-tuple $(\ell_1, \ell_2, \ldots, \ell_d)$ of its residue classes in $\mathbb{Z}/N_1\mathbb{Z}, \mathbb{Z}/N_2\mathbb{Z}, \ldots, \mathbb{Z}/N_d\mathbb{Z}$.

The proof of the CRT goes as follows. Let

$$Q_j = \frac{N}{N_j} = \prod_{i \neq j} N_i.$$

Since g.c.d. $(N_j, Q_j) = 1$ there exist integers n_j and q_j such that

$$n_j N_j + q_j Q_j = 1, \qquad j = 1, \ldots, d,$$

then the integer

$$\ell = \sum_{i=1}^{d} \ell_i q_i Q_i \bmod N$$

is the solution. Indeed,

$$\ell \equiv \ell_j q_j Q_j \bmod N_j$$

because all terms with $i \neq j$ contain N_j as a factor; and

$$q_j Q_j \equiv 1 \bmod N_j$$

by the defining relation for q_j.

It may be noted that

$$(q_i Q_i)(q_j Q_j) \equiv 0 \qquad \bmod N \text{ for } i \neq j,$$
$$(q_j Q_j)^2 \equiv q_j Q_j \qquad \bmod N, \ \ j = 1, \ldots, d,$$

so that the $q_j Q_j$ are mutually orthogonal *idempotents* in the ring $\mathbb{Z}/N\mathbb{Z}$, with properties formally similar to those of mutually orthogonal *projectors onto subspaces* in linear algebra. The analogy is exact, since by virtue of the CRT the ring $\mathbb{Z}/N\mathbb{Z}$ may be considered as the direct product

$$\mathbb{Z}/N_1\mathbb{Z} \times \mathbb{Z}/N_2\mathbb{Z} \times \ldots \times \mathbb{Z}/N_d\mathbb{Z}$$

via the two mutually inverse mappings:

(i) $\ell \longmapsto (\ell_1, \ell_2, \ldots, \ell_d)$ by $\ell \equiv \ell_j \bmod N_j$ for each j;

(ii) $(\ell_1, \ell_2, \ldots, \ell_d) \longmapsto \ell$ by $\ell = \sum_{i=1}^{d} \ell_i q_i Q_i \bmod N$.

The mapping defined by (ii) is sometimes called the 'CRT reconstruction' of ℓ from the ℓ_j.

These two mappings have the property of sending sums to sums and products to products, *i.e*:

(i) $\ell + \ell' \longmapsto (\ell_1 + \ell'_1, \ell_2 + \ell'_2, \ldots, \ell_d + \ell'_d)$
$\ell\ell' \longmapsto (\ell_1\ell'_1, \ell_2\ell'_2, \ldots, \ell_d\ell'_d)$

(ii) $(\ell_1 + \ell'_1, \ell_2 + \ell'_2, \ldots, \ell_d + \ell'_d) \longmapsto \ell + \ell'$
$(\ell_1\ell'_1, \ell_2\ell'_2, \ldots, \ell_d\ell'_d) \longmapsto \ell\ell'$

(the last proof requires using the properties of the idempotents $q_j Q_j$). This may be described formally by stating that the CRT establishes a *ring isomorphism*:

$$\mathbb{Z}/N\mathbb{Z} \cong (\mathbb{Z}/N_1\mathbb{Z}) \times \ldots \times (\mathbb{Z}/N_d\mathbb{Z}).$$

1.3.3.2.2.3. *The prime factor algorithm*

The CRT will now be used to factor the N-point DFT into a tensor product of d transforms, the jth of length N_j.

Let the indices k and k^* be subjected to the following mappings:

(i) $k \longmapsto (k_1, k_2, \ldots, k_d), k_j \in \mathbb{Z}/N_j\mathbb{Z}$, by $k_j \equiv k \bmod N_j$ for each j, with reconstruction formula

$$k = \sum_{i=1}^{d} k_i q_i Q_i \bmod N;$$

(ii) $k^* \longmapsto (k_1^*, k_2^*, \ldots, k_d^*), k_j^* \in \mathbb{Z}/N_j\mathbb{Z}$, by $k_j^* \equiv q_j k^* \bmod N_j$ for each j, with reconstruction formula

$$k^* = \sum_{i=1}^{d} k_i^* Q_i \bmod N.$$

Then

$$k^* k = \left(\sum_{i=1}^{d} k_i^* Q_i\right)\left(\sum_{j=1}^{d} k_j q_j Q_j\right) \bmod N$$
$$= \sum_{i,j=1}^{d} k_i^* k_j Q_i q_j Q_j \bmod N.$$

Cross terms with $i \neq j$ vanish since they contain all the factors of N, hence

$$k^* k = \sum_{j=1}^{d} q_j Q_j^2 k_j^* k_j \bmod N$$
$$= \sum_{j=1}^{d} (1 - n_j N_j) Q_j k_j^* k_j \bmod N.$$

Dividing by N, which may be written as $N_j Q_j$ for each j, yields

$$\frac{k^* k}{N} = \sum_{j=1}^{d} (1 - n_j N_j) \frac{Q_j}{N_j Q_j} k_j^* k_j \bmod 1$$
$$= \sum_{j=1}^{d} \left(\frac{1}{N_j} - n_j\right) k_j^* k_j \bmod 1,$$

and hence

$$\frac{k^* k}{N} \equiv \sum_{j=1}^{d} \frac{k_j^* k_j}{N_j} \bmod 1.$$

Therefore, by the multiplicative property of $e(.)$,

$$e\left(\frac{k^* k}{N}\right) \equiv \bigotimes_{j=1}^{d} e\left(\frac{k_j^* k_j}{N_j}\right).$$

Let $\mathbf{X} \in L(\mathbb{Z}/N\mathbb{Z})$ be described by a one-dimensional array $X(k)$ indexed by k. The index mapping (i) turns $\mathbf{X}$ into an element of $L(\mathbb{Z}/N_1\mathbb{Z} \times \ldots \times \mathbb{Z}/N_d\mathbb{Z})$ described by a d-dimensional array $X(k_1, \ldots, k_d)$; the latter may be transformed by $\bar{F}(N_1) \bigotimes \ldots \bigotimes \bar{F}(N_d)$ into a new array $X^*(k_1^*, k_2^*, \ldots, k_d^*)$. Finally, the one-dimensional array of results $X^*(k^*)$ will be obtained by reconstructing k^* according to (ii).

The prime factor algorithm, like the Cooley–Tukey algorithm, reindexes a 1D transform to turn it into d separate transforms, but the use of coprime factors and CRT index mapping leads to the further gain that *no twiddle factors* need to be applied between the successive transforms (see Good, 1971). This makes up for the cost of the added complexity of the CRT index mapping.

The natural factorization of N for the prime factor algorithm is thus its factorization into prime powers: $\bar{F}(N)$ is then the tensor product of separate transforms (one for each prime power factor $N_j = p_j^{\nu_j}$) whose results can be reassembled without twiddle factors. The separate factors p_j within each N_j must then be dealt with by another algorithm (*e.g.* Cooley–Tukey, which does require twiddle factors). Thus, the DFT on a prime number of points remains undecomposable.

1.3.3.2.3. *The Rader algorithm*

The previous two algorithms essentially reduce the calculation of the DFT on N points for N composite to the calculation of smaller DFTs on prime numbers of points, the latter remaining irreducible. However, Rader (1968) showed that the p-point DFT for p an odd

prime can itself be factored by invoking some extra arithmetic structure present in $\mathbb{Z}/p\mathbb{Z}$.

1.3.3.2.3.1. *N an odd prime*

The ring $\mathbb{Z}/p\mathbb{Z} = \{0, 1, 2, \ldots, p-1\}$ has the property that its $p-1$ non-zero elements, called *units*, form a *multiplicative group* $U(p)$. In particular, all units $r \in U(p)$ have a unique multiplicative inverse in $\mathbb{Z}/p\mathbb{Z}$, *i.e.* a unit $s \in U(p)$ such that $rs \equiv 1 \bmod p$. This endows $\mathbb{Z}/p\mathbb{Z}$ with the structure of a *finite field*.

Furthermore, $U(p)$ is a *cyclic* group, *i.e.* consists of the successive powers $g^m \bmod p$ of a generator g called a *primitive root mod p* (such a g may not be unique, but it always exists). For instance, for $p = 7$, $U(7) = \{1, 2, 3, 4, 5, 6\}$ is generated by $g = 3$, whose successive powers mod 7 are:

$$g^0 = 1, \quad g^1 = 3, \quad g^2 = 2, \quad g^3 = 6, \quad g^4 = 4, \quad g^5 = 5$$

[see Apostol (1976), Chapter 10].

The basis of Rader's algorithm is to bring to light a hidden regularity in the matrix $F(p)$ by permuting the basis vectors $\mathbf{u}_k$ and $\mathbf{v}_{k^*}$ of $L(\mathbb{Z}/p\mathbb{Z})$ as follows:

$$\begin{aligned}
\mathbf{u}'_0 &= \mathbf{u}_0 \\
\mathbf{u}'_m &= \mathbf{u}_k \quad \text{with } k = g^m, \qquad m = 1, \ldots, p-1; \\
\mathbf{v}'_0 &= \mathbf{v}_0 \\
\mathbf{v}'_{m^*} &= \mathbf{v}_{k^*} \quad \text{with } k^* = g^{m^*}, \quad m^* = 1, \ldots, p-1;
\end{aligned}$$

where g is a primitive root mod p.

With respect to these new bases, the matrix representing $\bar{F}(p)$ will have the following elements:

$$\begin{aligned}
\text{element } (0,0) &= 1 \\
\text{element } (0, m+1) &= 1 \quad \text{for all } m = 0, \ldots p-2, \\
\text{element } (m^*+1, 0) &= 1 \quad \text{for all } m^* = 0, \ldots, p-2, \\
\text{element } (m^*+1, m+1) &= e\left(\frac{k^* k}{p}\right) \\
&= e(g^{(m^*+m)/p}) \\
&\qquad \text{for all } m^* = 0, \ldots, p-2.
\end{aligned}$$

Thus the 'core' $\bar{C}(p)$ of matrix $\bar{F}(p)$, of size $(p-1) \times (p-1)$, formed by the elements with two non-zero indices, has a so-called *skew-circulant* structure because element (m^*, m) depends only on $m^* + m$. Simplification may now occur because multiplication by $\bar{C}(p)$ is closely related to a *cyclic convolution*. Introducing the notation $C(m) = e(g^{m/p})$ we may write the relation $\mathbf{Y}^* = \bar{F}(p)\mathbf{Y}$ in the permuted bases as

$$\begin{aligned}
Y^*(0) &= \sum_k Y(k) \\
Y^*(m^*+1) &= Y(0) + \sum_{m=0}^{p-2} C(m^*+m) Y(m+1) \\
&= Y(0) + \sum_{m=0}^{p-2} C(m^*-m) Z(m) \\
&= Y(0) + (\mathbf{C} * \mathbf{Z})(m^*), \quad m^* = 0, \ldots, p-2,
\end{aligned}$$

where $\mathbf{Z}$ is defined by $Z(m) = Y(p-m-2)$, $m = 0, \ldots, p-2$.

Thus $\mathbf{Y}^*$ may be obtained by cyclic convolution of $\mathbf{C}$ and $\mathbf{Z}$, which may for instance be calculated by

$$\mathbf{C} * \mathbf{Z} = F(p-1)[\bar{F}(p-1)[\mathbf{C}] \times \bar{F}(p-1)[\mathbf{Z}]],$$

where $\times$ denotes the component-wise multiplication of vectors. Since p is odd, $p-1$ is always divisible by 2 and may even be highly composite. In that case, factoring $\bar{F}(p-1)$ by means of the Cooley–Tukey or Good methods leads to an algorithm of complexity $p \log p$ rather than p^2 for $\bar{F}(p)$. An added bonus is that, because $g^{(p-1)/2} = -1$, the elements of $\bar{F}(p-1)[\mathbf{C}]$ can be shown to be either purely real or purely imaginary, which halves the number of real multiplications involved.

1.3.3.2.3.2. *N a power of an odd prime*

This idea was extended by Winograd (1976, 1978) to the treatment of prime powers $N = p^\nu$, using the cyclic structure of the multiplicative group of units $U(p^\nu)$. The latter consists of all those elements of $\mathbb{Z}/p^\nu\mathbb{Z}$ which are not divisible by p, and thus has $q_\nu = p^{\nu-1}(p-1)$ elements. It is cyclic, and there exist primitive roots g modulo p^ν such that

$$U(p^\nu) = \{1, g, g^2, g^3, \ldots, g^{q_\nu - 1}\}.$$

The $p^{\nu-1}$ elements divisible by p, which are divisors of zero, have to be treated separately just as 0 had to be treated separately for $N = p$.

When $k^* \notin U(p^\nu)$, then $k^* = pk_1^*$ with $k_1^* \in \mathbb{Z}/p^{\nu-1}\mathbb{Z}$. The results $X^*(pk_1^*)$ are p-decimated, hence can be obtained *via* the $p^{\nu-1}$-point DFT of the $p^{\nu-1}$-periodized data $\mathbf{Y}$:

$$X^*(pk_1^*) = \bar{F}(p^{\nu-1})[\mathbf{Y}](k_1^*)$$

with

$$Y(k_1) = \sum_{k_2 \in \mathbb{Z}/p\mathbb{Z}} X(k_1 + p^{\nu-1}k_2).$$

When $k^* \in U(p^\nu)$, then we may write

$$X^*(k^*) = X_0^*(k^*) + X_1^*(k^*),$$

where $\mathbf{X}_0^*$ contains the contributions from $k \notin U(p^\nu)$ and $\mathbf{X}_1^*$ those from $k \in U(p^\nu)$. By a converse of the previous calculation, $\mathbf{X}_0^*$ arises from p-decimated data $\mathbf{Z}$, hence is the $p^{\nu-1}$-periodization of the $p^{\nu-1}$-point DFT of these data:

$$X_0^*(p^{\nu-1}k_1^* + k_2^*) = \bar{F}(p^{\nu-1})[\mathbf{Z}](k_2^*)$$

with

$$Z(k_2) = X(pk_2), \qquad k_2 \in \mathbb{Z}/p^{\nu-1}\mathbb{Z}$$

(the $p^{\nu-1}$-periodicity follows implicity from the fact that the transform on the right-hand side is independent of $k_1^* \in \mathbb{Z}/p\mathbb{Z}$).

Finally, the contribution X_1^* from all $k \in U(p^\nu)$ may be calculated by reindexing by the powers of a primitive root g modulo p^ν, *i.e.* by writing

$$X_1^*(g^{m^*}) = \sum_{m=0}^{q_\nu - 1} X(g^m) e(g^{(m^*+m)/p^\nu})$$

then carrying out the multiplication by the skew-circulant matrix core as a convolution.

Thus the DFT of size p^ν may be reduced to two DFTs of size $p^{\nu-1}$ (dealing, respectively, with p-decimated results and p-decimated data) and a convolution of size $q_\nu = p^{\nu-1}(p-1)$. The latter may be 'diagonalized' into a multiplication by purely real or purely imaginary numbers (because $g^{(q_\nu/2)} = -1$) by two DFTs, whose factoring in turn leads to DFTs of size $p^{\nu-1}$ and $p-1$. This method, applied recursively, allows the complete decomposition of the DFT on p^ν points into arbitrarily small DFTs.

1.3.3.2.3.3. *N a power of 2*

When $N = 2^\nu$, the same method can be applied, except for a slight modification in the calculation of $\mathbf{X}_1^*$. There is no primitive root modulo 2^ν for $\nu > 2$: the group $U(2^\nu)$ is the direct product of *two* cyclic groups, the first (of order 2) generated by -1, the second (of order $N/4$) generated by 3 or 5. One then uses a representation

$$k = (-1)^{m_1} 5^{m_2}$$
$$k^* = (-1)^{m_1^*} 5^{m_2^*}$$

and the reindexed core matrix gives rise to a two-dimensional convolution. The latter may be carried out by means of two 2D DFTs on $2 \times (N/4)$ points.

1.3.3.2.4. *The Winograd algorithms*

The cyclic convolutions generated by Rader's multiplicative reindexing may be evaluated more economically than through DFTs if they are re-examined within a new algebraic setting, namely the theory of congruence classes of polynomials [see, for instance, Blahut (1985), Chapter 2; Schroeder (1986), Chapter 24].

The set, denoted $\mathbb{K}[X]$, of polynomials in one variable with coefficients in a given field $\mathbb{K}$ has many of the formal properties of the set $\mathbb{Z}$ of rational integers: it is a *ring* with no zero divisors and has a *Euclidean algorithm* on which a theory of divisibility can be built.

Given a polynomial $P(z)$, then for every $W(z)$ there exist unique polynomials $Q(z)$ and $R(z)$ such that

$$W(z) = P(z)Q(z) + R(z)$$

and

$$\text{degree } (R) \; < \; \text{degree } (P).$$

$R(z)$ is called the *residue* of $H(z)$ modulo $P(z)$. Two polynomials $H_1(z)$ and $H_2(z)$ having the same residue modulo $P(z)$ are said to be *congruent* modulo $P(z)$, which is denoted by

$$H_1(z) \equiv H_2(z) \bmod P(z).$$

If $H(z) \equiv 0 \bmod P(z)$, $H(z)$ is said to be *divisible* by $P(z)$. If $H(z)$ only has divisors of degree zero in $\mathbb{K}[X]$, it is said to be *irreducible over* $\mathbb{K}$ (this notion depends on $\mathbb{K}$). Irreducible polynomials play in $\mathbb{K}[X]$ a role analogous to that of prime numbers in $\mathbb{Z}$, and any polynomial over $\mathbb{K}$ has an essentially unique factorization as a product of irreducible polynomials.

There exists a *Chinese remainder theorem* (CRT) for polynomials. Let $P(z) = P_1(z) \ldots P_d(z)$ be factored into a product of pairwise coprime polynomials [*i.e.* $P_i(z)$ and $P_j(z)$ have no common factor for $i \neq j$]. Then the system of congruence equations

$$H(z) \equiv H_j(z) \bmod P_j(z), \quad j = 1, \ldots, d,$$

has a unique solution $H(z)$ modulo $P(z)$. This solution may be constructed by a procedure similar to that used for integers. Let

$$Q_j(z) = P(z)/P_j(z) = \prod_{i \neq j} P_i(z).$$

Then P_j and Q_j are coprime, and the Euclidean algorithm may be used to obtain polynomials $p_j(z)$ and $q_j(z)$ such that

$$p_j(z)P_j(z) + q_j(z)Q_j(z) = 1.$$

With $S_i(z) = q_i(z)Q_i(z)$, the polynomial

$$H(z) = \sum_{i=1}^{d} S_i(z)H_i(z) \bmod P(z)$$

is easily shown to be the desired solution.

As with integers, it can be shown that the 1:1 correspondence between $H(z)$ and $H_j(z)$ sends sums to sums and products to products, *i.e.* establishes a *ring isomorphism*:

$$\mathbb{K}[X] \bmod P \cong (\mathbb{K}[X] \bmod P_1) \times \ldots \times (\mathbb{K}[X] \bmod P_d).$$

These results will now be applied to the efficient calculation of cyclic convolutions. Let $\mathbf{U} = (u_0, u_1, \ldots, u_{N-1})$ and $\mathbf{V} = (v_0, v_1, \ldots, v_{N-1})$ be two vectors of length N, and let $\mathbf{W} = (w_0, w_1, \ldots, w_{N-1})$ be obtained by cyclic convolution of $\mathbf{U}$ and $\mathbf{V}$:

$$w_n = \sum_{m=0}^{N-1} u_m v_{n-m}, \quad n = 0, \ldots, N-1.$$

The very simple but crucial result is that this cyclic convolution may be carried out by *polynomial multiplication modulo* $(z^N - 1)$: if

$$U(z) = \sum_{l=0}^{N-1} u_l z^l$$
$$V(z) = \sum_{m=0}^{N-1} v_m z^m$$
$$W(z) = \sum_{n=0}^{N-1} w_n z^n$$

then the above relation is equivalent to

$$W(z) \equiv U(z)V(z) \bmod (z^N - 1).$$

Now the polynomial $z^N - 1$ can be *factored* over the field of rational numbers into irreducible factors called *cyclotomic polynomials*: if d is the number of divisors of N, including 1 and N, then

$$z^N - 1 = \prod_{i=1}^{d} P_i(z),$$

where the cyclotomics $P_i(z)$ are well known (Nussbaumer, 1981; Schroeder, 1986, Chapter 22). We may now invoke the CRT, and exploit the ring isomorphism it establishes to simplify the calculation of $W(z)$ from $U(z)$ and $V(z)$ as follows:

(i) compute the d residual polynomials

$$U_i(z) \equiv U(z) \bmod P_i(z), \quad i = 1, \ldots, d,$$
$$V_i(z) \equiv V(z) \bmod P_i(z), \quad i = 1, \ldots, d;$$

(ii) compute the d polynomial products

$$W_i(z) \equiv U_i(z)V_i(z) \bmod P_i(z), \quad i = 1, \ldots, d;$$

(iii) use the CRT reconstruction formula just proved to recover $W(z)$ from the $W_i(z)$:

$$W(z) \equiv \sum_{i=1}^{d} S_i(z)W_i(z) \bmod (z^N - 1).$$

When N is not too large, *i.e.* for 'short cyclic convolutions', the $P_i(z)$ are very simple, with coefficients 0 or ± 1, so that (i) only involves a small number of additions. Furthermore, special techniques have been developed to multiply general polynomials modulo cyclotomic polynomials, thus helping keep the number of multiplications in (ii) and (iii) to a minimum. As a result, cyclic convolutions can be calculated rapidly when N is sufficiently composite.

It will be recalled that Rader's multiplicative indexing often gives rise to cyclic convolutions of length $p - 1$ for p an odd prime. Since $p - 1$ is highly composite for all $p \leq 50$ other than 23 and 47, these cyclic convolutions can be performed more efficiently by the above procedure than by DFT.

These combined algorithms are due to Winograd (1977, 1978, 1980), and are known collectively as 'Winograd small FFT algorithms'. Winograd also showed that they can be thought of as bringing the DFT matrix $\mathbf{F}$ to the following 'normal form':

$$\mathbf{F} = \mathbf{CBA},$$

where

$\mathbf{A}$ is an integer matrix with entries 0, ± 1, defining the 'pre-additions',

B is a diagonal matrix of multiplications,

C is a matrix with entries $0, \pm 1, \pm i$, defining the 'post-additions'.

The elements on the diagonal of **B** can be shown to be either real or pure imaginary, by the same argument as in Section 1.3.3.2.3.1. Matrices **A** and **C** may be rectangular rather than square, so that intermediate results may require extra storage space.

1.3.3.3. *Multidimensional algorithms*

From an algorithmic point of view, the distinction between one-dimensional (1D) and multidimensional DFTs is somewhat blurred by the fact that some factoring techniques turn a 1D transform into a multidimensional one. The distinction made here, however, is a practical one and is based on the dimensionality of the indexing sets for data and results. This section will therefore be concerned with the problem of factoring the DFT when the *indexing sets* for the input data and output results are multidimensional.

1.3.3.3.1. *The method of successive one-dimensional transforms*

The DFT was defined in Section 1.3.2.7.4 in an n-dimensional setting and it was shown that when the decimation matrix **N** is diagonal, say $\mathbf{N} = \mathrm{diag}(N^{(1)}, N^{(2)}, \ldots, N^{(n)})$, then $\bar{F}(N)$ has a tensor product structure:

$$\bar{F}(\mathbf{N}) = \bar{F}(N^{(1)}) \otimes \bar{F}(N^{(2)}) \otimes \ldots \otimes \bar{F}(N^{(n)}).$$

This may be rewritten as follows:

$$\begin{aligned}\bar{F}(\mathbf{N}) = {} & [\bar{F}(N^{(1)}) \otimes I_{N^{(2)}} \otimes \ldots \otimes I_{N^{(n)}}] \\ & \times [I_{N^{(1)}} \otimes \bar{F}(N^{(2)}) \otimes \ldots \otimes I_{N^{(n)}}] \\ & \times \ldots \\ & \times [I_{N^{(1)}} \otimes I_{N^{(2)}} \otimes \ldots \otimes \bar{F}(N^{(n)}],\end{aligned}$$

where the I's are identity matrices and $\times$ denotes ordinary matrix multiplication. The matrix within each bracket represents a one-dimensional DFT along one of the n dimensions, the other dimensions being left untransformed. As these matrices commute, the order in which the successive 1D DFTs are performed is immaterial.

This is the most straightforward method for building an n-dimensional algorithm from existing 1D algorithms. It is known in crystallography under the name of 'Beevers–Lipson factorization' (Section 1.3.4.3.1), and in signal processing as the 'row–column method'.

1.3.3.3.2. *Multidimensional factorization*

Substantial reductions in the arithmetic cost, as well as gains in flexibility, can be obtained if the factoring of the DFT is carried out in several dimensions simultaneously. The presentation given here is a generalization of that of Mersereau & Speake (1981), using the abstract setting established independently by Auslander, Tolimieri & Winograd (1982).

Let us return to the general n-dimensional setting of Section 1.3.2.7.4, where the DFT was defined for an arbitrary decimation matrix **N** by the formulae (where $|\mathbf{N}|$ denotes $|\det \mathbf{N}|$):

$$F(\mathbf{N}): \quad X(\mathbf{k}) = \frac{1}{|\mathbf{N}|} \sum_{\mathbf{k}^*} X^*(\mathbf{k}^*) e[-\mathbf{k}^* \cdot (\mathbf{N}^{-1}\mathbf{k})]$$

$$\bar{F}(\mathbf{N}): \quad X^*(\mathbf{k}^*) = \sum_{\mathbf{k}} X(\mathbf{k}) e[\mathbf{k}^* \cdot (\mathbf{N}^{-1}\mathbf{k})]$$

with

$$\mathbf{k} \in \mathbb{Z}^n/\mathbf{N}\mathbb{Z}^n, \quad \mathbf{k}^* \in \mathbb{Z}^n/\mathbf{N}^T\mathbb{Z}^n.$$

1.3.3.3.2.1. *Multidimensional Cooley–Tukey factorization*

Let us now assume that this decimation can be factored into d successive decimations, *i.e.* that

$$\mathbf{N} = \mathbf{N}_1\mathbf{N}_2 \ldots \mathbf{N}_{d-1}\mathbf{N}_d$$

and hence

$$\mathbf{N}^T = \mathbf{N}_d^T\mathbf{N}_{d-1}^T \ldots \mathbf{N}_2^T\mathbf{N}_1^T.$$

Then the coset decomposition formulae corresponding to these successive decimations (Section 1.3.2.7.1) can be combined as follows:

$$\begin{aligned}\mathbb{Z}^n &= \bigcup_{\mathbf{k}_1} (\mathbf{k}_1 + \mathbf{N}_1\mathbb{Z}^n) \\ &= \bigcup_{\mathbf{k}_1} \left\{ \mathbf{k}_1 + \mathbf{N}_1 \left[\bigcup_{\mathbf{k}_2} (\mathbf{k}_2 + \mathbf{N}_2\mathbb{Z}^n) \right] \right\} \\ &= \ldots \\ &= \bigcup_{\mathbf{k}_1} \ldots \bigcup_{\mathbf{k}_d} (\mathbf{k}_1 + \mathbf{N}_1\mathbf{k}_2 + \ldots + \mathbf{N}_1\mathbf{N}_2 \times \ldots \times \mathbf{N}_{d-1}\mathbf{k}_d + \mathbf{N}\mathbb{Z}^n)\end{aligned}$$

with $\mathbf{k}_j \in \mathbb{Z}^n/\mathbf{N}_j\mathbb{Z}^n$. Therefore, any $\mathbf{k} \in \mathbb{Z}/\mathbf{N}\mathbb{Z}^n$ may be written uniquely as

$$\mathbf{k} = \mathbf{k}_1 + \mathbf{N}_1\mathbf{k}_2 + \ldots + \mathbf{N}_1\mathbf{N}_2 \times \ldots \times \mathbf{N}_{d-1}\mathbf{k}_d.$$

Similarly:

$$\begin{aligned}\mathbb{Z}^n &= \bigcup_{\mathbf{k}_d^*} (\mathbf{k}_d^* + \mathbf{N}_d^T\mathbb{Z}^n) \\ &= \ldots \\ &= \bigcup_{\mathbf{k}_d^*} \ldots \bigcup_{\mathbf{k}_1^*} (\mathbf{k}_d^* + \mathbf{N}_d^T\mathbf{k}_{d-1}^* + \ldots + \mathbf{N}_d^T \times \ldots \times \mathbf{N}_2^T\mathbf{k}_1^* \\ &\quad + \mathbf{N}^T\mathbb{Z}^n)\end{aligned}$$

so that any $\mathbf{k}^* \in \mathbb{Z}^n/\mathbf{N}^T\mathbb{Z}^n$ may be written uniquely as

$$\mathbf{k}^* = \mathbf{k}_d^* + \mathbf{N}_d^T\mathbf{k}_{d-1}^* + \ldots + \mathbf{N}_d^T \times \ldots \times \mathbf{N}_2^T\mathbf{k}_1^*$$

with $\mathbf{k}_j^* \in \mathbb{Z}^n/\mathbf{N}_j^T\mathbb{Z}^n$. These decompositions are the vector analogues of the multi-radix number representation systems used in the Cooley–Tukey factorization.

We may then write the definition of $\bar{F}(\mathbf{N})$ with $d = 2$ factors as

$$\begin{aligned}X^*(\mathbf{k}_2^* + \mathbf{N}_2^T\mathbf{k}_1^*) = {} & \sum_{\mathbf{k}_1} \sum_{\mathbf{k}_2} X(\mathbf{k}_1 + \mathbf{N}_1\mathbf{k}_2) \\ & \times e[(\mathbf{k}_2^{*T} + \mathbf{k}_1^{*T}\mathbf{N}_2)\mathbf{N}_2^{-1}\mathbf{N}_1^{-1}(\mathbf{k}_1 + \mathbf{N}_1\mathbf{k}_2)].\end{aligned}$$

The argument of $e(-)$ may be expanded as

$$\mathbf{k}_2^* \cdot (\mathbf{N}^{-1}\mathbf{k}_1) + \mathbf{k}_1^* \cdot (\mathbf{N}_1^{-1}\mathbf{k}_1) + \mathbf{k}_2^* \cdot (\mathbf{N}_2^{-1}\mathbf{k}_2) + \mathbf{k}_1^* \cdot \mathbf{k}_2.$$

The first summand may be recognized as a twiddle factor, the second and third as the kernels of $\bar{F}(\mathbf{N}_1)$ and $\bar{F}(\mathbf{N}_2)$, respectively, while the fourth is an integer which may be dropped. We are thus led to a 'vector-radix' version of the Cooley–Tukey algorithm, in which the successive decimations may be introduced in all n dimensions simultaneously by general integer matrices. The computation may be decomposed into five stages analogous to those of the one-dimensional algorithm of Section 1.3.3.2.1:

(i) form the $|\mathbf{N}_1|$ vectors $\mathbf{Y}_{\mathbf{k}_1}$ of shape $\mathbf{N}_2$ by

$$Y_{\mathbf{k}_1}(\mathbf{k}_2) = X(\mathbf{k}_1 + \mathbf{N}_1\mathbf{k}_2), \quad \mathbf{k}_1 \in \mathbb{Z}^n/\mathbf{N}_1\mathbb{Z}^n, \quad \mathbf{k}_2 \in \mathbb{Z}^n/\mathbf{N}_2\mathbb{Z}^n;$$

references