Fast generalized DFTs for all finite groups

Chris Umans
Computing & Mathematical Sciences
California Institute of Technology
umans@cs.caltech.edu

***Abstract*—For any finite group G, we give an arithmetic algorithm to compute generalized Discrete Fourier Transforms (DFTs) with respect to G, using $O(|G|^{\omega/2+\epsilon})$ operations, for any $\epsilon > 0$. Here, ω is the exponent of matrix multiplication.**

***Keywords*-Discrete Fourier Transform, finite group, algorithm**

I. Introduction

For a finite group G, let $\mathrm{Irr}(G)$ denote a complete set of irreducible representations of G. A *generalized DFT with respect to* G is a map from a group algebra element $\alpha \in \mathbb{C}[G]$ (which is a vector of $|G|$ complex numbers), to the following linear combination of irreducible representations:

$$\sum_{g\in G} \alpha_g \bigoplus_{\rho\in\mathrm{Irr}(G)} \rho(g).$$

This is the fundamental linear operation that maps the standard basis for the group algebra $\mathbb{C}[G]$ to the Fourier basis of irreducible representations of group G. It has applications in data analysis [1], machine learning [2], optimization [3], as a component in other algorithms (including fast operations on polynomials and in the Cohn-Umans matrix multiplication algorithms), and as the basis for quantum algorithms for problems entailing a Hidden Subgroup Problem [4].

This paper gives algorithms that compute generalized DFTs with respect to any finite group G, and any chosen bases for the ρ. We typically speak of the complexity of computing a generalized DFT map in the (non-uniform) arithmetic circuit model and do not concern ourselves with *finding* the irreducible representations. The trivial algorithm thus requires $O(|G|^2)$ operations, since one can simply sum up $|G|$ block-diagonal matrices, each with $|G|$ entries in the blocks.

Fast algorithms for the DFT with respect to cyclic groups are well-known and are attributed to Cooley and Tukey in 1965 [5], although the ideas likely date to Gauss. Beth in 1984 [6], together with Clausen [7], initiated the study of generalized DFTs, the "generalized" terminology signalling that the underlying group may be any group. A central goal since that time has been to obtain fast algorithms for generalized DFTs with respect to arbitrary underlying groups. One may hope for "nearly-linear" time algorithms, meaning that they use a number of operations that is upper-bounded by $c_\epsilon|G|^{1+\epsilon}$ for universal constants c_ϵ and arbitrary $\epsilon > 0$. Such "exponent one" algorithms are known for certain families of groups: abelian groups, supersolvable groups [8], and symmetric and alternating groups [7]. Algorithms for generalized DFTs manipulate matrices, so it is not surprising that they often require a number of operations that depends on ω, the exponent of matrix multiplication. Thus we view algorithms that achieve exponent one conditioned on $\omega = 2$ as being "nearly as good" as unconditional exponent one algorithms. Such algorithms are known for solvable groups [6], [9], and with the recent breakthrough of [10], for linear groups; these algorithms achieve exponent $\omega/2$.

In this paper we realize the main goal of the area, obtaining exponent $\omega/2$ for all finite groups G. The previous best exponent that applies to all finite groups was obtained by [10]; it depends in a somewhat complicated way on ω, but it is at best $\sqrt{2}$ (when $\omega = 2$); our exponent beats the one obtained by [10] for every ω between 2 and 3. Before [10], the best known exponent was $1+\omega/4$ (which is at best $3/2$ when $\omega = 2$), and this dates back to the original work of Beth and Clausen.

A. Past and related work

A good description of past work in this area can be found in Section 13.5 of [11]. The first algorithm generalizing beyond the abelian case is due to Beth in 1984 [6]; this algorithm is described in Section III-A in a form often credited jointly to Beth and Clausen. Three other milestones are the $O(|G|\log|G|)$ algorithm for supersolvable groups due to Baum [8], the $O(|G|\log^3|G|)$ algorithm for the symmetric group due to Clausen [7] (see also [12] for a recent improvement), and the $O(|G|^{\omega/2+\epsilon})$ algorithms for linear groups obtained by Hsu and Umans, which are described in Section III-B. Wreath products were studied by Rockmore [13] who obtained exponent one algorithms in certain cases.

In the 1990s, Maslen, Rockmore, and coauthors developed the so-called "separation of variables" approach [14], which relies on non-trivial decompositions along chains of subgroups via *Bratteli diagrams* and detailed knowledge of the representation theory of the underlying groups. There is a rather large body of literature on this approach and it has been applied to a wide variety of group algebras and

Supported by NSF grant CCF-1815607 and a Simons Foundation Investigator grant.

2575-8454/19/$31.00 ©2019 IEEE
DOI 10.1109/FOCS.2019.00052

more general algebraic objects. For a fuller description of this approach and the results obtained, the reader is referred to the surveys [4], [15], and the most recent paper in this line of work [16].

II. Preliminaries

Throughout this paper we will use the phrase

"generalized DFTs w. r. t. G can be computed using $O(|G|^{\alpha+\epsilon})$ operations, for all $\epsilon > 0$"

where G is a finite group and $\alpha \geq 1$ is a real number. We mean by this that there are *universal* constants c_ϵ independent of the group G under consideration so that for each $\epsilon > 0$, the operation count is at most $c_\epsilon |G|^{\alpha+\epsilon}$. Such an algorithm will be referred to as an "exponent α" algorithm. This comports with the precise definition of the exponent of matrix multiplication, ω: that there are universal constants b_ϵ for which $n \times n$ matrix multiplication can be performed using at most $b_\epsilon n^{\omega+\epsilon}$ operations, for each $\epsilon > 0$. Indeed we will often report our algorithms' operation counts in terms of ω. In such cases matrix multiplication is always used as a black box, so, for example, an operation count of $O(|G|^{\omega/2})$ should be interpreted to mean: if one uses a fast matrix multiplication algorithm with exponent α (which may range from 2 to 3), then the operation count is $O(|G|^{\alpha/2})$. In particular, in real implementations, one might well use standard matrix multiplication and plug in 3 for ω in the operation count bound.

We use $\mathrm{Irr}(G)$ to denote the complete set of irreducible representations of G being used for the DFT at hand. In the presentation to follow, we assume the underlying field is $\mathbb{C}$; however our algorithms work over any field $\mathbb{F}_{p^k}$ whose characteristic p does not divide the order of the group, and for which k is sufficiently large for $\mathbb{F}_{p^k}$ to represent a complete set of irreducibles.

We use I_n to denote the $n \times n$ identity matrix. The following is an important general observation (see, e.g., Lemma 4.3.1 in [17]):

Proposition 1. *If A is an $n_1 \times n_2$ matrix, B is an $n_2 \times n_3$ matrix, and C is an $n_3 \times n_4$ matrix, then the entries of the product matrix ABC are exactly the entries of the vector obtained by multiplying $A \otimes C^T$ (which is an $n_1 n_4 \times n_2 n_3$ matrix) by B viewed as an $n_2 n_3$-vector, which is denoted vec(B).*

A. Basic representation theory

A *representation* of group G is a homomorphism ρ from G into the group of invertible $d \times d$ matrices. Representation ρ naturally specifies an action of G on $\mathbb{C}^d$; representation ρ is thus said to have *dimension* $\dim(\rho) = d$. A representation is *irreducible* if the action on $\mathbb{C}^d$ has no G-invariant subspace. Two representations of the same dimension d, ρ_1 and ρ_2, are *equivalent* (written $\rho_1 \cong \rho_2$) if they are the same up to a change of basis; i.e., $\rho_1(g) = T\rho_2(g)T^{-1}$ for some invertible $d \times d$ matrix T. The classical Maschke's Theorem implies that every representation ρ_0 of G *breaks up* into the direct sum of irreducible representations; i.e. there is an invertible matrix T and a multiset $S \subseteq \mathrm{Irr}(G)$, for which

$$T\rho_0(g)T^{-1} = \bigoplus_{\rho \in S} \rho(g).$$

Given a subgroup $H \subseteq G$ one can obtain from any representation $\rho \in \mathrm{Irr}(G)$ a representation $\mathrm{Res}^G_H(\rho)$ (the *restriction* of ρ to H), which is a representation of H, simply by restricting the domain of ρ to H. One can also obtain from any representation $\sigma \in \mathrm{Irr}(H)$, a representation of G called the *induced* representation $\mathrm{Ind}^G_H(\rho)$, which has dimension $\dim(\sigma)|G|/|H|$. We will not need to work directly with induced representations, but we will use a fundamental fact called *Frobenius reciprocity*. Given $\rho \in \mathrm{Irr}(G)$ and $\sigma \in \mathrm{Irr}(H)$, Frobenius reciprocity states that the number of times σ appears in the restriction $\mathrm{Res}^G_H(\rho)$ equals the number of times ρ appears in the induced representation $\mathrm{Ind}^G_H(\sigma)$.

A basic fact is that $\sum_{\rho \in \mathrm{Irr}(G)} \dim(\rho)^2 = |G|$, which implies that for all $\rho \in \mathrm{Irr}(G)$, we have $\dim(\rho) \leq |G|^{1/2}$. This can be used to prove the following inequality, which we use repeatedly:

Proposition 2. *For any real number $\alpha \geq 2$, we have*

$$\sum_{\rho \in \mathrm{Irr}(G)} \dim(\rho)^\alpha \leq |G|^{\alpha/2}.$$

Proof: Set $\rho_{\max}$ to be an irrep of largest dimension. We have

$$\begin{aligned}\sum_{\rho \in \mathrm{Irr}(G)} \dim(\rho)^\alpha &\leq \dim(\rho_{\max})^{\alpha-2} \sum_{\rho \in \mathrm{Irr}(G)} \dim(\rho)^2 \\ &= \dim(\rho_{\max})^{\alpha-2}|G| \leq |G|^{\alpha/2},\end{aligned}$$

where the last inequality used the fact that $\dim(\rho_{\max}) \leq |G|^{1/2}$. ■

B. Basic Clifford theory

Clifford theory describes the way the irreducible representations of a group H break up when restricted to a *normal* subgroup N, which is a particularly well-structured and well-understood scenario.

Elements of H act on the set $\mathrm{Irr}(N)$ as follows:

$$(h \cdot \lambda)(n) = \lambda(hnh^{-1}),$$

for $\lambda \in \mathrm{Irr}(N)$. Let $\mathcal{O}_1, \ldots, \mathcal{O}_\ell$ be the orbits of this H-action on $\mathrm{Irr}(N)$. Clifford theory states for each $\sigma \in \mathrm{Irr}(H)$, there is a positive integer e_σ and an index i_σ for which the restriction $\mathrm{Res}^H_N(\sigma)$ is equivalent to

$$e_\sigma \bigoplus_{\lambda \in \mathcal{O}_{i_\sigma}} \lambda.$$

In particular, this implies that all $\lambda \in \mathrm{Irr}(N)$ that occur in the restriction have the same dimension, d_σ, and multiplicity, e_σ, and that $\dim(\sigma) = d_\sigma e_\sigma |\mathcal{O}_{i_\sigma}|$.

We can also define the following subsets, which partition $\mathrm{Irr}(H)$:

$$\begin{aligned} S_\ell &= \{\sigma \in \mathrm{Irr}(H) : \text{the irreps in } \mathcal{O}_\ell \text{ occur in } \sigma\} \\ &= \{\sigma \in \mathrm{Irr}(H) : i_\sigma = \ell\}. \end{aligned}$$

We will need the following proposition:

Proposition 3. *For a finite group H and normal subgroup N, and sets S_ℓ as defined above, the following holds for each ℓ:*

$$\sum_{\sigma \in S_\ell} \dim(\sigma) e_\sigma / d_\sigma = |H/N|.$$

Proof: Fix $\lambda \in \mathcal{O}_\ell$, and note that the induced representation $\mathrm{Ind}_N^H(\lambda)$ has dimension $\dim(\lambda)|H/N|$. Let $m_{\sigma,\lambda}$ be the number of times $\sigma \in \mathrm{Irr}(H)$ occurs in $\mathrm{Ind}_N^H(\lambda)$. Then we have

$$\sum_{\sigma \in \mathrm{Irr}(H)} \dim(\sigma) m_{\sigma,\lambda} = \dim(\lambda)|H/N|.$$

By Frobenius reciprocity, $m_{\sigma,\lambda}$ equals the number times λ occurs in $\mathrm{Res}_N^H(\sigma)$. Thus the summand $\dim(\sigma) m_{\sigma,\lambda}$ equals $\dim(\sigma) e_\sigma$, whenever $m_{\sigma,\lambda} \neq 0$ (and zero otherwise). The proposition follows. ∎

C. Generalized DFTs and inverse generalized DFTs

We assume by default that we are computing generalized DFTs with respect to an arbitrary chosen basis for each $\rho \in \mathrm{Irr}(G)$. Sometimes we need to refer to the special basis in the following definition:

Definition 4. *Let H be a subgroup of G. An H-*adapted basis *is a basis for each $\rho \in \mathrm{Irr}(G)$, so that the restriction of ρ to H respects the direct sum decomposition into irreps of H.*

In concrete terms, this implies that for each $\rho \in \mathrm{Irr}(G)$, while for general $g \in G$, $\rho(g)$ is a $\dim(\rho) \times \dim(\rho)$ matrix, for $g \in H$, $\rho(g)$ is a block-diagonal matrix with block sizes coming from the set $\{\dim(\sigma) : \sigma \in \mathrm{Irr}(H)\}$. An H-adapted basis always exists.

A general trick that we will rely on is that if one can compute generalized DFTs with respect to G for an input α supported on a subset $S \subseteq G$, then with an additional multiplicative factor of roughly $|G|/|S|$, one can compute generalized DFTs with respect to G.

Theorem 5. *Fix a finite group G and a subset $S \subseteq G$, and suppose a generalized DFT with respect to G can be computed in m operations, for inputs α supported on S. Then generalized DFTs with respect to G can be computed using*

$$O(m + |G|^{\omega/2+\epsilon}) \cdot \frac{|G| \log |G|}{|S|}$$

operations, for any $\epsilon > 0$.

Proof: First observe that by multiplying by $\oplus_{\rho \in \mathrm{Irr}(G)} \rho(g)$ we can compute a generalized DFT supported on Sg, for an additive extra cost of

$$\sum_{\rho \in \mathrm{Irr}(G)} O(\dim(\rho)^{\omega+\epsilon})$$

operations, for all $\epsilon > 0$, and by applying Proposition 2 with $\alpha = \omega + \epsilon$ this is at most $O(|G|^{\omega/2+\epsilon})$. A probabalistic argument shows that $|G| \log |G|/|S|$ different translations g of S suffice to cover G, so we need only repeat the DFT supported on Sg translated by each such g, and sum the resulting DFTs. ∎

The *inverse generalized DFT* maps a collection of matrices $M^\sigma \in \mathbb{C}^{\dim(\sigma) \times \dim(\sigma)}$, one for each $\sigma \in \mathrm{Irr}(G)$, to the vector α for which

$$\sum_{g \in G} \alpha_g \bigoplus_{\sigma \in \mathrm{Irr}(G)} \rho(G) = \bigoplus_{\sigma \in \mathrm{Irr}(G)} M^\sigma.$$

In the arithmetic circuit model, the inverse DFT can be computed efficiently if the DFT can:

Theorem 6 (Baum, Clausen; Cor. 13.40 in [11])**.** *Fix a generalized DFT with respect to finite group G and suppose it can be computed in m operations. Then the inverse DFT with respect to G (and the same basis), can be computed in at most $m + |G|$ operations.*

D. Main technical ideas

Here we highlight three key technical ideas that go into the main result.

1) Structure in an H-DFT when H has a normal subgroup: In general, an H-DFT

$$\sum_{h \in H} \alpha_h \bigoplus_{\sigma \in \mathrm{Irr}(H)} \sigma(h)$$

is a block-diagonal matrix with $\sum_{\sigma \in \mathrm{Irr}(H)} \dim(\sigma)^2 = |H|$ non-zero entries, or "degrees of freedom". If H has a subgroup N with coset representatives X, the H-DFT can be equivalently written

$$\sum_{n \in N} \left(\underbrace{\sum_{x \in X} \alpha_{xn} \bigoplus_{\sigma \in \mathrm{Irr}(H)} \sigma(x)}_{M_n} \right) \cdot \bigoplus_{\sigma \in \mathrm{Irr}(H)} \sigma(n).$$

We show in Theorem 11 that *if N is normal*, then matrix M_n can be taken to have special structure well beyond the block-diagonal structure of an H-DFT: various entries can be made to repeat in a prescribed pattern, in the same way for all n. Then, just as we describe a block-diagonal matrix as having a number of "degrees of freedom" equal to the number of entries in the blocks, we can describe the M_n as having a number of "degrees of freedom" equal to the number of free entries, and in the structure we uncover in this paper, this number is the information-theoretic optimal, $|H|/|N|$. This structure is *accessible* in the sense that it can be efficiently obtained from an H-DFT, by performing a number of inverse N-DFTs, and it is the key to overcoming the bottleneck in the previous best result [10].

2) Efficient matrix multiplication for certain block-structured matrices: In order to make use of the above structured matrices in our recursive algorithm, we need to be able to multiply them with a vector efficiently. The following situation arises: we have a matrix with several "big" blocks along the diagonal, with each big block itself being a block-diagonal matrix. The big blocks have the same number of entries but incompatible structure, and the entries in each big block are repeated in each other big block, in a pattern we can choose. For example, two of the big blocks might look like the block-diagonal matrices in the top row of Figure 2. It is straightforward to multiply such a matrix with a vector in time proportional to the number of free entries *times the number of big blocks*. We devise a way to multiply such a matrix with a vector in time proportional to only the number of free entries, paying only a logarithmic price as overhead (see Section III-C1 and Lemma 14).

3) Triple subgroup structure in every finite group: One of the challenges in designing an algorithm computing generalized DFTs with respect to an arbitrary finite group G is that the algorithm can only exploit structure that can be found in *every* finite group. Beyond the Sylow Theorems, there is very little to work with. Past work made use of Lev's Theorem, which states that every finite group (other than a cyclic group) has a moderately large subgroup, and [10] made use of the Classification Theorem to prove that every finite group (other than a p-group) has two proper subgroups H and K whose product HK nearly covers the entire group. However $H \cap K$ may be quite large, which limits the usefulness of this decomposition. Our main structural result on groups (Theorem 16) strengthens the decomposition of [10] to prove that every finite group has a normal subgroup N (possibly trivial) for which G/N is either cyclic of prime order, or has proper subgroups H, K with $H \cap K = \{1\}$ and whose product HK nearly covers the entire quotient group. In other words, after quotient-ing by a normal subgroup, every group is either cyclic of prime order, or "almost" a so-called *Zappa-Szép product*. This structural result seems natural and potentially useful beyond the application in this paper.

III. General strategy: reduction to subgroups

One way to organize the main algorithmic ideas in the quest for a fast DFT for all finite groups is according to the subgroup structure they exploit. The algorithms themselves are recursive, with the main content of the algorithm being the reduction to smaller instances: DFTs over subgroups of the original group. When aiming for generalized DFTs for all finite groups, such a reduction is paired with a group-theoretic structural result, which guarantees the existence of certain subgroups that are used by the reduction.

In the exposition below, it is helpful to assume that $\omega = 2$ and seek an "exponent 1" algorithm under this assumption (in general, the exponent achieved will be a function of ω, and in our main result this function is $\omega/2$). By the term *overhead* we mean the extra multiplicative factor in the operation count of the reduction, beyond the nearly-linear operation count that would be necessary for an exponent 1 algorithm.

A. The single subgroup reduction

The seminal Beth-Clausen algorithm reduces computing a DFT over a group G to computing several DFTs over a subgroup H of G. We call this the "single subgroup reduction". Roughly speaking, the overhead in this reduction is proportional to the index of H in G. The companion structural result is Lev's Theorem [18], which shows that every finite group G (except cyclic of prime order which can be handled separately) has a subgroup of order at least $\sqrt{G}$ (and this is tight, hence the overhead is $\sqrt{|G|}$ in the worst case). As noted in the introduction, this reduction together with Lev's Theorem implies exponent $3/2$ (assuming $\omega = 2$) for all finite groups.

Here is a more detailed description, together with results we will need later. Let H be a subgroup of G and let X be a set of distinct coset representatives. We first compute several H-DFTs, one for each $x \in X$:

$$s_x = \sum_{h \in H} \alpha_{hx} \bigoplus_{\sigma \in \mathrm{Irr}(H)} \sigma(h)$$

and by using an H*-adapted* basis (Definition 4), we can lift each s_x to

$$\overline{s_x} = \sum_{h \in H} \alpha_{hx} \bigoplus_{\rho \in \mathrm{Irr}(G)} \rho(h)$$

by just copying entries (which is free of cost in the arithmetic model). Then to complete the DFT we need to compute

$$\sum_{x \in X} \overline{s_x} \bigoplus_{\rho \in \mathrm{Irr}(G)} \rho(x).$$

The $\rho(x)$ factors in the equation are often called "twiddle factors" when G is abelian. Generically, this final computation requires an overhead proportional to $|X| = [G : H]$, even when just considering the outermost summation. See Corollary 4 in [19] for the details to complete this sketch, yielding the following:

Theorem 7 (single subgroup reduction)**.** *Let G be a finite group and let H be a subgroup. Then we can compute a generalized DFT with respect to G at a cost of $[G : H]$ many H-DFTs plus $O([G : H]|G|^{\omega/2+\epsilon})$ operations, for all $\epsilon > 0$.*

In the special case that H is normal in G and G/H is cyclic of prime order, the overhead of $[G : H]$ can be avoided, by using knowledge about the way representations $\sigma \in \mathrm{Irr}(H)$ extend to $\rho \in \mathrm{Irr}(G)$. This insight is the basis for the Beth-Clausen algorithm for solvable groups. We need it here to handle the case of G/H cyclic of prime order, which is the single exceptional case not handled by our main reduction. The following theorem can be inferred from the proof of Theorem 7.7 in Clausen and Baum's monograph [9]:

Theorem 8 (Clausen, Baum [9])**.** *Let H be a normal subgroup of G with prime index p. We can compute a generalized DFT with respect to G and an H-adapted basis, at a cost of p many H-DFTs plus*

$$O(p \log p) \cdot \sum_{\sigma \in \mathrm{Irr}(H)} \dim(\sigma)^{\omega+\epsilon}$$

operations, for all $\epsilon > 0$.

For our purposes the following slightly coarser bound suffices, which accommodates an arbitary basis change (hence obviating the need for an H-adapted basis):

Corollary 9. *Let H be a normal subgroup of G with prime index p. Generalized DFTs with respect to G can be computed at a cost of p many H-DFTs plus $O(|G|^{\omega/2+\epsilon})$ operations, for all $\epsilon > 0$.*

Proof: Applying Proposition 2 to Theorem 8 with $\alpha = \omega+\epsilon$ yields an operation count of $O(p \log p)|H|^{\omega+\epsilon/2}$, which is at most $O(|G|^{\omega/2+\epsilon})$. Performing an arbitary basis change costs

$$\sum_{\rho \in \mathrm{Irr}(G)} O(\dim(\rho)^{\omega+\epsilon})$$

operations which is again at most $O(|G|^{\omega/2+\epsilon})$ by Proposition 2. ∎

B. The double subgroup reduction

Recently, Hsu and Umans proposed a "double subgroup reduction" [10] which reduces computing a DFT over a group G to computing several DFTs over two subgroups, H and K. This reduction is especially effective for linear groups (see [10]). Roughly speaking, the overhead in this reduction is proportional to $|G|/|HK|$ and $|H \cap K|$. The companion structural result shows that every finite group G (except p-groups which can be handled separately) has two proper subgroups H and K for which $|G|/|HK|$ is negligible. However, $|H \cap K|$ might still be large, which is the one thing standing in the way of deriving an "exponent $\omega/2$" algorithm from this reduction.

To illustrate the bottleneck in this reduction, we describe it in more detail. Let H, K be subgroups of G and assume $|G|/|HK|$ is negligible. We first compute an intermediate representation

$$\sum_{g=hk \in HK} \alpha_g \bigoplus_{\substack{\sigma \in \mathrm{Irr}(H)\\ \tau \in \mathrm{Irr}(K)}} \sigma(h) \otimes \tau(k)$$

in two steps (and then lift it to a G-DFT). The first of the two steps is to compute at most $[G : H]$ many H-DFTs, yielding, for each $k \in K' \subseteq K$ (where K' is a set of distinct coset representatives of H in G):

$$s_k = \sum_{h \in H} \alpha_{hk} \bigoplus_{\sigma \in \mathrm{Irr}(H)} \sigma(h).$$

The second step is as follows: for each *entry* of the block-diagonal matrix s_k, we use this entry (as k varies) as the data for a K-DFT. There are $\sum_{\sigma \in \mathrm{Irr}(H)} \dim(\sigma)^2 = |H|$ such entries in general. Thus the second step entails $|H|$ many K-DFTs, and this represents the key bottleneck. Note that when $|G|/|HK|$ is negligible, $|H||K|$ is approximately $|G||H \cap K|$, and this explains the overhead of roughly $|H \cap K|$ which prevents obtaining an "exponent $\omega/2$" algorithm from this reduction. For completeness we record the main theorem of [19] here:

Theorem 10 (Theorem 12 in [19])**.** *Let G be a finite group and let H, K be subgroups. Then we can compute generalized DFTs with respect to G at the cost of $|H|$ many K-DFTS, $|K|$ many H-DFTs, plus*

$$O(|G|^{\omega/2+\epsilon} + (|H||K|)^{\omega/2+\epsilon})$$

operations, all repeated $O(\frac{|G| \log |G|}{|HK|})$ times, for all $\epsilon > 0$.

Our main innovation, described in the next section, is a way to overcome the bottleneck. When $H \cap K = N$ is a normal subgroup of G, we are able to rewrite each s_k as a sum of $|N|$ matrices with special structure: effectively, there are only $|H/N|$ many non-zero "entries" for which we need to compute a K-DFT, and as we will show, this exactly removes the overhead factor.

C. The triple subgroup reduction

In this section we give our main new result. We devise a "triple subgroup reduction" which reduces computing a DFT over G to computing several DFTs over two subgroups, H and K, *and* several inverse DFTs over the intersection $N = H \cap K$, when N is normal in G. Roughly speaking, the overhead is proportional to $|G|/|HK|$. The companion structural result (Theorem 16) shows that for every finite group G, if N is a maximal normal subgroup in G then (except for the case of $|G/N|$ cyclic of prime order, which can be handled separately) there exist two proper subgroups H and K with $H \cap K = N$, such that $|G|/|HK|$ is negligible. This is the key to the claimed exponent $\omega/2$ algorithm.

Let H be a group with normal subgroup N. The main technical theorem shows how to rewrite the output of an

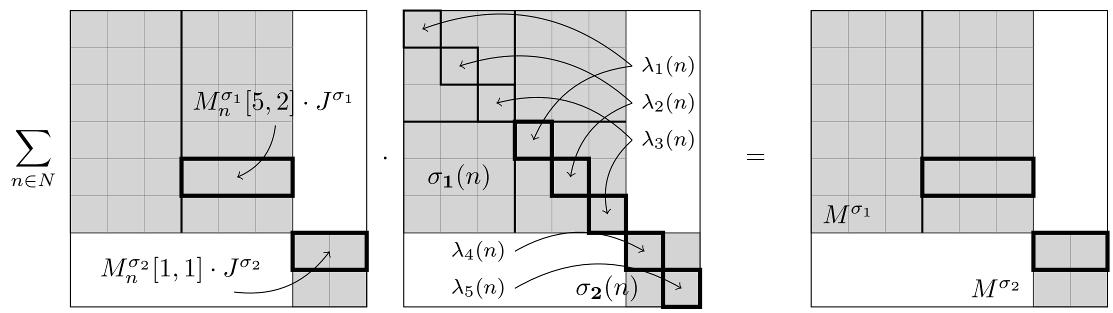

Figure 1. Illustration of the proof of Theorem 11. In this example $\mathrm{Irr}(H)=\{\sigma_1,\sigma_2\}$, $\mathrm{Irr}(N)=\{\lambda_1,\lambda_2,\lambda_3,\lambda_4,\lambda_5\}$; the orbits are $\mathcal{O}_1=\{\lambda_1,\lambda_2,\lambda_3\}$ and $\mathcal{O}_2=\{\lambda_4,\lambda_5\}$; $S_1=\{\sigma_1\}$ and $S_2=\{\sigma_2\}$; and the multiplicities are $e_{\sigma_1}=2$ and $e_{\sigma_2}=1$. In the figure, we highlight the parts of the matrices that give rise to the system of equations solved with a single inverse N-DFT, corresponding to the value $a=f_1(\sigma_1,5,2)=f_2(\sigma_2,1,1)$. This inverse N-DFT with the highlighted blocks of M^{σ_1} and M^{σ_2} as input data yields the scalars $M_n^{\sigma_1}[5,2]=M_n^{\sigma_2}[1,1]$ that satisfy the simultaneous equations.

H-DFT as the sum of $|N|$ matrices each of which only has "$|H/N|$ degrees of freedom". In the following theorem we adopt the notation introduced in Section II-B; as a reminder: d_σ is the dimension of the N-irreps occurring in the restriction $\mathrm{Res}^H_N(\sigma)$, e_σ is the multiplicity, and $\mathcal{O}_\ell$ are the orbits of the H-action on $\mathrm{Irr}(N)$, which are used to define the sets S_ℓ which partition $\mathrm{Irr}(H)$.

Theorem 11. *Let H be a group and N a normal subgroup. For every*

$$M=\bigoplus_{\sigma\in\mathrm{Irr}(H)} M^\sigma \in \bigoplus_{\sigma\in\mathrm{Irr}(H)} \mathbb{C}^{\dim(\sigma)\times\dim(\sigma)},$$

the following holds with respect to an N-adapted basis: there exist matrices $M_n^\sigma\in\mathbb{C}^{\dim(\sigma)/d_\sigma\times e_\sigma}$ for which

$$\sum_{n\in N}(M_n^\sigma\otimes J^\sigma)\cdot\sigma(n)=M^\sigma,$$

where J^σ is the $d_\sigma\times\dim(\sigma)/e_\sigma$ matrix $(I_{d_\sigma}|I_{d_\sigma}|\cdots|I_{d_\sigma})$. Moreover, given injective functions f_ℓ from $\{(\sigma,i,j):\sigma\in S_\ell, i\in[\dim(\sigma)/d_\sigma], j\in[e_\sigma]\}$ to $[r]$, the M_n^σ can be taken to satisfy

$$f_\ell(\sigma,i,j)=f_{\ell'}(\sigma',i',j') \;\Rightarrow\; \forall n\; M_n^\sigma[i,j]=M_n^{\sigma'}[i',j'],$$

and these matrices M_n^σ can be obtained from M by computing r inverse N-DFTs.

One should think of the functions f_ℓ as labeling the entries of the M_n^σ matrices for the σ in a given S_ℓ. This labeling is then used to ensure that entries of M_n^σ with $\sigma\in S_\ell$ and the entries of $M_n^{\sigma'}$ with $\sigma'\in S_{\ell'}$ are equal, if they have the same labels. In Section III-C1 we will show how to choose this labeling so that the final "lifting" step of our algorithm can be efficiently computed. For now, we note that Proposition 3 implies that there *exist* labelings f_ℓ with $r=|H/N|$, and indeed our actual choice of f_ℓ in Section III-C1 will have $r=O(|H/N|\log|H/N|)$, which is not much larger.

Proof: Fix $\sigma\in\mathrm{Irr}(H)$, and recall that there is a unique S_ℓ containing σ. Since we are using an N-adapted basis, $\sigma(n)$ has the form

$$I_{e_\sigma}\otimes\bigoplus_{\lambda\in\mathcal{O}_\ell}\lambda(n),$$

and thus

$$\sum_{n\in N}(M_n^\sigma\otimes J^\sigma)\cdot\sigma(n) = \sum_{n\in N} M_n^\sigma\otimes(\lambda_1(n)|\lambda_2(n)|\cdots|\lambda_{|\mathcal{O}_\ell|}(n)) \quad (1)$$

where $\lambda_1,\ldots,\lambda_{|\mathcal{O}_\ell|}$ is an enumeration of $\mathcal{O}_\ell$. Since these are pairwise inequivalent irreps, the span of

$$\{(\lambda_1(n)|\lambda_2(n)|\cdots|\lambda_{|\mathcal{O}_\ell|}(n)):n\in N\}$$

is the full matrix algebra $\mathbb{C}^{d_\sigma\times\dim(\sigma)/e_\sigma}$. Hence we can choose the M_n^σ so that expression (1) equals an arbitrary $M^\sigma\in\mathbb{C}^{\dim(\sigma)\times\dim(\sigma)}$.

In particular, for each σ, the (i,j) entries of the M_n^σ should satisfy

$$\sum_{n\in N}M_n^\sigma[i,j]\begin{pmatrix}\lambda_1(n)\\ \lambda_2(n)\\ \vdots\\ \lambda_v(n)\end{pmatrix}=\begin{pmatrix}M^\sigma[i,jv]\\ M^\sigma[i,jv+1]\\ \vdots\\ M^\sigma[i,jv+v-1]\end{pmatrix} \quad (2)$$

where $v=|\mathcal{O}_\ell|$ and M^σ occurring on the right-hand-side is blocked into $d_\sigma\times d_\sigma$ submatrices and indexed accordingly. Thus the values of a given entry of M_n^σ as n ranges over N, can be found in an inverse N-DFT with the appropriate blocks of M^σ as input data.

Observe however that in general, $\mathcal{O}_\ell$ is a *proper* subset of $\mathrm{Irr}(H)$, and hence the aforementioned inverse N-DFT is underdetermined; for example Equation (2) remains satisfied if we require $\sum_{n\in N}M_n^\sigma[i,j]\lambda(n)=0$ for all $\lambda\in\mathrm{Irr}(H)\setminus\mathcal{O}_\ell$.

Indeed, we can *simultaneously* solve Equation (2) with respect to several $\sigma\in\mathrm{Irr}(H)$ via a single inverse N-DFT, provided the associated orbits $\mathcal{O}_{i_\sigma}$ are different. To prove the "moreover" part of the theorem statement, then, we set up the following system of equations, for a given $a \in [r]$: for each ℓ for which $f_\ell(\sigma, i, j) = a$ we *simultaneously* require that Equation (2) holds with respect to σ, i, j (and note these are determined by a since f_ℓ is injective). Since the S_ℓ partition $\mathrm{Irr}(H)$, selecting at most one σ from each S_ℓ results in a system that mentions each $\lambda \in \mathrm{Irr}(N)$ at most once. Hence a single inverse N-DFT solves this system of equations. See Figure 1. We do this once for each $a \in [r]$, to produce the matrices M_n^σ from the original M, using r inverse N-DFTs. ∎

1) Choosing the labelings f_ℓ: To make use of Theorem 11, we need to define injective functions f_ℓ from

$$\{(\sigma, i, j) : \sigma \in S_\ell, i \in [\dim(\sigma)/d_\sigma], j \in [e_\sigma]\}$$

to $[r]$. We identify the domain of f_ℓ with the entries of a block-diagonal matrix, with rectangular blocks of size $\dim(\sigma)/d_\sigma \times e_\sigma$, as σ ranges over S_ℓ. Recall that by Proposition 3, the total number of entries in these blocks is $|H/N|$.

We will describe functions f_ℓ associating the entries of a block-diagonal matrix of this format (which depends on ℓ) with a *target* block-diagonal matrix whose format is fixed as follows:

$$\begin{array}{lll}
2 \cdot |H/N| & \text{blocks of size} & 1 \times 1 \\
\lceil 2 \cdot |H/N|/4 \rceil & \text{blocks of size} & 2 \times 2 \\
\lceil 2 \cdot |H/N|/16 \rceil & \text{blocks of size} & 4 \times 4 \\
 & \vdots & \\
\lceil 2 \cdot |H/N|/2^{2i} \rceil & \text{blocks of size} & 2^i \times 2^i \\
 & \vdots & \\
2 & \text{blocks of size} & \left(2^{\lceil \log_2 |H/N| \rceil}\right)^2
\end{array}$$

Note that the number of entries of this target matrix is $O(|H/N| \log |H/N|)$, and this will be our r. The association specifying the map f_ℓ is quite simple: we take one column at a time of the source block-diagonal matrix, and if it has height w, we associate it, top-aligned, with the next-available column among the blocks of size $2^i \times 2^i$, for the i such that $2^i/2 < w \le 2^i$. See Figure 2. Since there can be at most $|H/N|/w < 2|H/N|/2^i$ columns of height w in the source matrix (which has $|H/N|$ entries in total), and the target block-diagonal matrix has at least $2 \cdot |H/N|/2^i$ columns of height 2^i, this association is possible.

We will use these mappings when applying Theorem 11 to facilitate an efficient "lift" from an intermediate representation to the final G-DFT. The key benefit of the mappings is that they allow us to combine several matrix-vector products with incompatible matrix formats into one, as illustrated in Figure 2. In order to be able to speak precisely about this combined object, we make the following definition:

Definition 12 (parent matrix)**.** *Given a partition of* $\mathrm{Irr}(H)$ *into sets S_ℓ, matrices A^σ with dimensions* $\dim(\sigma)/d_\sigma \times e_\sigma$ *(one for each $\sigma \in$* $\mathrm{Irr}(H)$*), and functions f_ℓ as above, satisfying*

$$f_\ell(\sigma, i, j) = f_{\ell'}(\sigma', i', j') \quad \Rightarrow \quad A^\sigma[i, j] = A^{\sigma'}[i', j'],$$

define the parent matrix *of the A^σ to be the matrix with the format of the target matrix above, and with entry (x, y) equal to the value of $A^\sigma[i, j]$ if there exists ℓ for which $f_\ell(\sigma, i, j) = (x, y)$, and zero otherwise.*

See Figure 3 for an example parent matrix.

2) Computing the intermediate representation: We are at the point now where we can compute the intermediate representation, which we then lift to the final G-DFT in Lemma 14, making critical use of the just-described labelings f_ℓ. The setup is as follows: H and K are proper subgroups of group G, and $H \cap K = N$ is normal in G. Let X be a system of distinct coset representatives of N in H and let Y be a system of distinct coset representatives of N in K. Thus $H = XN$ and $K = NY$. Note that $HK = XNY$ with uniqueness of expression.

When applying the triple subgroup reduction in our final result, it will happen that

$$\frac{|G|}{|HK|} = \frac{|G||N|}{|H||K|}$$

is negligible, and notice that in this case, if H-DFTs, K-DFTs, and N-DFTS have nearly-linear algorithms, then indeed the cost of applying the next lemma is nearly-linear in $|G|$ as desired.

Lemma 13. *With $|Y|$ many H-DFTs, $O(|H/N| \log |H/N|) \cdot |Y|$ many inverse N-DFTs, and $O(|H/N| \log |H/N|)$ many K-DFTs, we can compute, from $\alpha \in \mathbb{C}[G]$ supported on HK, the following expression:*

$$\sum_{n \in N} \sum_{y \in Y} \bigoplus_{\tau \in \mathrm{Irr}(K)} P_{n,y} \otimes \tau(ny)^T \tag{3}$$

where $P_{n,y}$ is the parent matrix of the matrices $\{M_{n,y}^\sigma : \sigma \in$ $\mathrm{Irr}(H)\}$*, and for each σ, y, the $M_{n,y}^\sigma$ satisfy (with respect to an N-adapted basis for* $\mathrm{Irr}(H)$*):*

$$\sum_{n \in N} (M_{n,y}^\sigma \otimes J^\sigma) \sigma(n) = \sum_{h \in H} \alpha_{hy} \sigma(h). \tag{4}$$

where J^σ is the $\dim(\sigma)/e_\sigma \times d_\sigma$ *matrix $(I_{d_\sigma} | I_{d_\sigma} | \cdots | I_{d_\sigma})$ as in Theorem 11.*

Expression (3) arises in Equation (9) in the next section after manipulating the expression for a G-DFT supported on $HK = HY$, and it is the "input" to Lemma 14 which efficiently lifts it to a G-DFT.

Proof: First, compute for each $y \in Y$ and $\sigma \in \mathrm{Irr}(H)$ the matrices

$$M_y^\sigma = \sum_{h \in H} \alpha_{hy} \sigma(h),$$

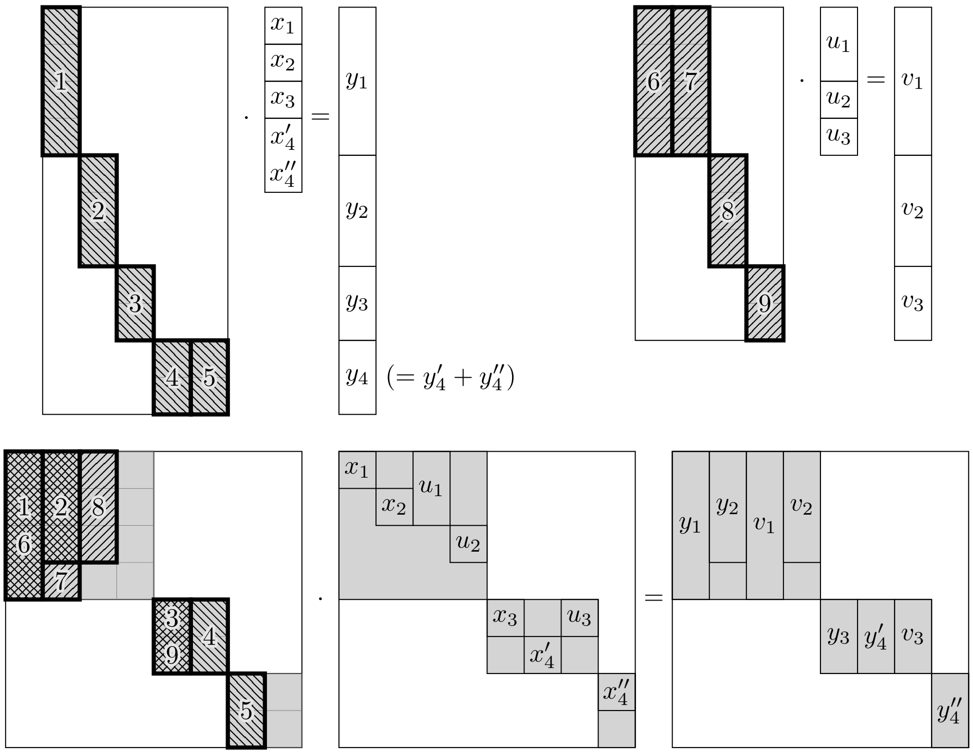

Figure 2. Example illustrating how the f_ℓ functions are defined and used. The numbered columns of the block-diagonal matrix in the upper-left are associated to the columns of the target block-diagonal matrix on the bottom-left in the manner described in Section III-C1. The numbered columns of the block-diagonal matrix in the upper-right are also associated by the same procedure, and the figure shows these two associations superimposed on each other. We see that the two matrix-vector multiplications at the top can be combined into the single matrix product on the bottom, provided that similarly labeled entries of the two source matrices are guaranteed to contain identical values. Unlabeled cells of the middle-bottom matrix contain zeros. Note that in the bottom-right matrix each segment of the original vectors y and v may be padded up to twice its original length (but not more), and it may be repeated up to twice and summed (as y'_4 and y''_4 are) if the columns of the associated block are mapped to two different blocks in the target matrix. More than two repetitions are not possible because the source blocks all have at most as many columns as rows.

using $|Y|$ different H-DFTs. Next, apply Theorem 11, once for each y, to the matrices

$$\bigoplus_{\sigma \in \mathrm{Irr}(H)} M_y^\sigma \in \bigoplus_{\sigma \in \mathrm{Irr}(H)} \mathbb{C}^{\dim(\sigma)\times\dim(\sigma)},$$

together with the labelings f_ℓ from Section III-C1, to obtain matrices $M_{n,y}^\sigma \in \mathbb{C}^{\dim(\sigma)/d_\sigma \times e_\sigma}$ for which

$$\sum_{n\in N} (M_{n,y}^\sigma \otimes J^\sigma)\sigma(n) = M_y^\sigma,$$

at a cost of $O(|H/N|\log|H/N|)\cdot|Y|$ many inverse N-DFTs. Note that these $M_{n,y}^\sigma$ satisfy Equation (4). Let $P_{n,y}$ be the parent matrix of the matrices $\{M_{n,y}^\sigma : \sigma \in \mathrm{Irr}(H)\}$.

For each (i,j), the vector β with $\beta[ny] = P_{n,y}[i,j]$ is an element of $\mathbb{C}[K]$ and we perfom a K-DFT on it; this entails computing at most $O(|H/N|\log|H/N|)$ different K-DFTs because this is the number of entries in the blocks of the block-diagonal matrices $P_{n,y}$. At this point we hold, in the aggregate, all of the entries of Expression (3) in the statement of the lemma, and the proof is complete. ■

3) Lifting to a G-DFT: In this section we show how to efficiently lift the intermediate representation, Expression (3) computed via Lemma 13, to a G-DFT. We continue with the notation of the previous section.

Let $\mathrm{Irr}^*(H)$ denote the *multiset* of irreps of H that occur in the restrictions of the irreps of G to H (with the correct multiplicities), and similarly let $\mathrm{Irr}^*(K)$ denote the *multiset* of irreps of K that occur in the restrictions of the irreps of G to K. Let S and T be the change of basis matrices that satisfy:

$$S\left(\bigoplus_{\sigma\in\mathrm{Irr}^*(H)} \sigma(h)\right)S^{-1} = \bigoplus_{\rho\in\mathrm{Irr}(G)} \rho(h) \quad \forall h \in H$$
$$T\left(\bigoplus_{\tau\in\mathrm{Irr}^*(K)} \tau(k)\right)T^{-1} = \bigoplus_{\rho\in\mathrm{Irr}(G)} \rho(k) \quad \forall k \in K.$$

We further specify that S should be with respect to an N-adapted basis for $\mathrm{Irr}(H)$.

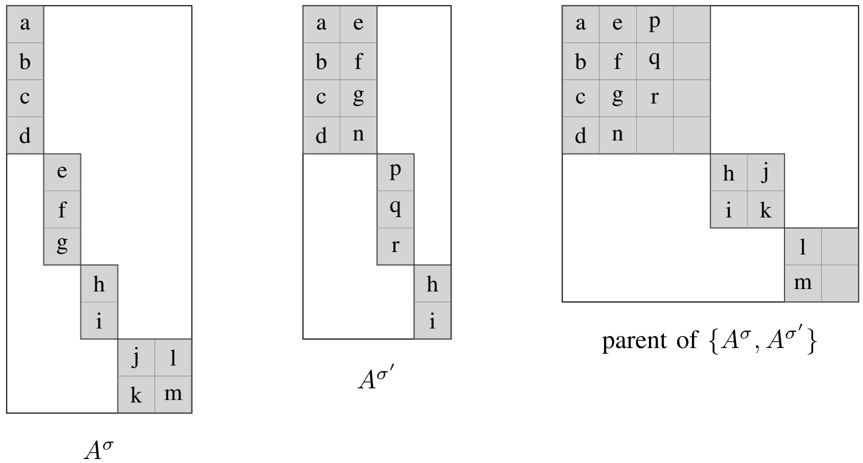

Figure 3. An example parent matrix. Unlabeled entries are zero. Empty blocks in the parent matrix are not pictured.

Notice that for $n \in N = H \cap K$, we have:

$$S\left(\bigoplus_{\sigma\in\mathrm{Irr}^*(H)} \sigma(n)\right) S^{-1} = T\left(\bigoplus_{\tau\in\mathrm{Irr}^*(K)} \tau(n)\right) T^{-1},$$

or equivalently

$$\left(\bigoplus_{\sigma\in\mathrm{Irr}^*(H)} \sigma(n)\right) S^{-1}T = S^{-1}T\left(\bigoplus_{\tau\in\mathrm{Irr}^*(K)} \tau(n)\right), \quad (5)$$

a fact we will use shortly.

A G-DFT with input α supported on $HY = HK$ is the expression:

$$\begin{aligned}
&\sum_{\substack{h\in H\\ y\in Y}} \alpha_{hy} \bigoplus_{\rho\in\mathrm{Irr}(G)} \rho(hy)\\
&= \sum_{y\in Y}\left(\sum_{h\in H}\alpha_{hy}\bigoplus_{\rho\in\mathrm{Irr}(G)}\rho(h)\right)\cdot\left(\bigoplus_{\rho\in\mathrm{Irr}(G)}\rho(y)\right)\\
&= \sum_{y\in Y} S\left(\sum_{h\in H}\alpha_{hy}\bigoplus_{\sigma\in\mathrm{Irr}^*(H)}\sigma(h)\right) S^{-1}T\left(\bigoplus_{\tau\in\mathrm{Irr}^*(K)}\tau(y)\right)T^{-1}
\end{aligned}$$

Now for each $y \in Y$, the left-most parenthesized expression is an H-DFT, with certain blocks repeated. Set $R = S^{-1}T$. By Equation (4) in the statement of Lemma 13, each such expression can be rewritten in terms of matrices $M^\sigma_{n,y}$, yielding:

$$\begin{aligned}
&\sum_{\substack{h\in H\\ y\in Y}} \alpha_{hy} \bigoplus_{\rho\in\mathrm{Irr}(G)} \rho(hy) =\\
&\sum_{\substack{y\in Y\\ n\in N}} S\left(\bigoplus_{\sigma\in\mathrm{Irr}^*(H)} (M^\sigma_{n,y}\otimes J^\sigma)\sigma(n)\right) R\left(\bigoplus_{\tau\in\mathrm{Irr}^*(K)}\tau(y)\right)T^{-1}\\
&= \sum_{\substack{y\in Y\\ n\in N}} \underbrace{S\left(\bigoplus_{\sigma\in\mathrm{Irr}^*(H)} (M^\sigma_{n,y}\otimes J^\sigma)\right) R\left(\bigoplus_{\tau\in\mathrm{Irr}^*(K)}\tau(ny)\right)T^{-1}}_{(*)}
\end{aligned} \quad (6)$$

where the last line invoked Equation (5) to move $\sigma(n)$ past $R = S^{-1}T$.

We now focus on Expression $(*)$. By Proposition 1 we can express Expression $(*)$ as

$$\left(\bigoplus_{\substack{\sigma\in\mathrm{Irr}^*(H)\\ \tau\in\mathrm{Irr}^*(K)}} \left((M^\sigma_{n,y}\otimes J^\sigma)\otimes\tau(ny)^T\right)\right)\cdot \mathrm{vec}(R) = \mathrm{vec}(*). \quad (7)$$

We next apply two types of simplifications to the block-diagonal matrix on the left. In each, we observe that equalities among blocks allow us to simplify that block-diagonal matrix, at the expense of arranging portions of vec(R) and vec$(*)$ into block-diagonal matrices and summing certain entries. The first such observation is that computing

$$\left(\begin{array}{c|c} A & \\ \hline & A\end{array}\right)\cdot\left(\frac{x_1}{x_2}\right) = \left(\frac{y_1}{y_2}\right)$$

is equivalent to computing $A\cdot(x_1|x_2) = (y_1|y_2)$. The second observation is that computing

$$(A|A) \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = y$$

is equivalent to computing $A \cdot (x_1 + x_2) = y$.

Using the first observation we can thus simplify Equation (7) to:

$$\left(\bigoplus_{\substack{\sigma \in \mathrm{Irr}(H) \\ \tau \in \mathrm{Irr}(K)}} \left((M^\sigma_{n,y} \otimes J^\sigma) \otimes \tau(ny)^T \right) \right) \cdot X_0 = Y_0,$$

where X_0 is a block-diagonal matrix whose entries coincide with the entries of R. Next, we notice that

$$J^\sigma = I_{d_\sigma} \otimes (1, 1, \ldots 1).$$

The first observation then allows us to simplify Equation (7) futher to:

$$\left(\bigoplus_{\substack{\sigma \in \mathrm{Irr}(H) \\ \tau \in \mathrm{Irr}(K)}} \left((M^\sigma_{n,y} \otimes (1, 1, \ldots 1)) \otimes \tau(ny)^T \right) \right) \cdot X_1 = Y_1$$

where again the entries of X_1 coincide with the entries of R, and the second observation allows us to simplify to:

$$\left(\bigoplus_{\substack{\sigma \in \mathrm{Irr}(H) \\ \tau \in \mathrm{Irr}(K)}} M^\sigma_{n,y} \otimes \tau(ny)^T \right) \cdot X_2 = Y_2, \qquad (8)$$

where now X_2 is a block-diagonal matrix whose entries are sums of entries of R.

As in the statement of Lemma 13, for each n, y, let $P_{n,y}$ be the parent matrix of the matrices $\{M^\sigma_{n,y} : \sigma \in \mathrm{Irr}(H)\}$. We can rewrite Equation (8) as

$$\left(\bigoplus_{\tau \in \mathrm{Irr}(K)} P_{n,y} \otimes \tau(ny)^T \right) \cdot X_3 = Y_3, \qquad (9)$$

where X_3 is again a block-diagonal matrix whose entries are sums of entries of R.

The square blocks of the block-diagonal matrix

$$\left(\bigoplus_{\tau \in \mathrm{Irr}(K)} P_{n,y} \otimes \tau(ny)^T \right)$$

have dimensions a_i with the property that

$$\sum_i a_i^2 = O(|H/N| \log |H/N|) \cdot |K|,$$

using our earlier accounting for the block sizes of a parent matrix, together with the fact that $\sum_{\tau \in \mathrm{Irr}(K)} \dim(\tau)^2 = |K|$. Each $a_i \times a_i$ block is multiplied by an $a_i \times w_i$ block of X_3, to yield an $a_i \times w_i$ block of the product matrix Y_3. We now argue that the w_i satisfy $\sum_i a_i w_i \leq 4|G|$. Each of the two transformations applied to obtain block-diagonal matrices Y_0, Y_1 and then Y_2 preserve the number of entries of the result matrix; these matrices therefore have $|G|$ entries in the blocks since Y_0 does. The final transformation results in a block-diagonal matrix Y_3 which may have *more* entries than $|G|$, but this number can be larger by only a factor of four, as illustrated in Figure 2. This is because each column of a block of Y_2 may need to be padded to at most twice its original length, and repeated up to two times (and no more, because the blocks of the $M^\sigma_{n,y}$ have no more columns than rows, and thus can spill over at most two blocks in the parent matrix). Thus the number of entries in the blocks of Y_3 which equals $\sum_i a_i w_i$, is at most $4|G|$ as stated.

We conclude that the block-matrix multiplication in Equation (9) can be performed efficiently as summarized in the following lemma.

Lemma 14. *The map from*

$$\sum_{n \in N} \sum_{y \in Y} \bigoplus_{\tau \in \mathrm{Irr}(K)} P_{n,y} \otimes \tau(ny)^T$$

as computed from input α supported on $HY = HK$ in Lemma 13, to a G-DFT, can be computed at a cost of $O(|G|^{\omega/2+\epsilon})$ operations, for all $\epsilon > 0$.

Proof: We describe how to map a summand $\bigoplus_{\tau \in \mathrm{Irr}(K)} P_{n,y} \otimes \tau(ny)^T$ to the corresponding summand of Expression (6). This map will be *linear* and will not depend on n, y, so we apply it once to the entire sum computed by Lemma 13, to obtain Expression (6), which is the promised G-DFT.

We need to perform matrix multiplications of format $\langle a_i, a_i, w_i \rangle$, and we know that $\sum_i a_i^2 = O(|H/N| \log |H/N|) \cdot |K| = L$ and $\sum_i a_i w_i \leq 4|G|$. The cost of such a multiplication is at most $\max(O(a_i^{\omega+\epsilon}), O(a_i^{\omega-1+\epsilon} w_i))$ for all $\epsilon > 0$. Replacing the maximum with a sum, and letting $a_{\max} = \max_i a_i$, we obtain an upper bound on the number of operations of

$$\begin{aligned} \sum_i O(a_i^{\omega+\epsilon}) &+ O(a_i^{\omega-1+\epsilon} w_i) \\ &= O(a_{\max}^{\omega-2+\epsilon}) \sum_i a_i^2 + a_i w_i \\ &\leq L^{(\omega-2+\epsilon)/2} \cdot (L + 4|G|). \qquad (10) \end{aligned}$$

We need to pre-multiply by S and post-multiply by T^{-1} to obtain a summand of Expression (6). Both S and T^{-1} are block-diagonal with one block for each $\rho \in \mathrm{Irr}(G)$, with dimension $\dim(\rho)$. Thus the cost of this final pre- and post-multiplication is

$$\sum_{\rho \in \mathrm{Irr}(G)} O(\dim(\rho)^{\omega+\epsilon})$$

which is at most $O(|G|^{\omega/2+\epsilon})$ by Proposition 2 with $\alpha = \omega + \epsilon$. The theorem follows from the fact that $|H||K|/|N| \leq$

$|G|$, and thus Expression (10) is also upper-bounded by $O(|G|^{\omega/2+\epsilon})$ (absorbing logarithmic terms into $|G|^{\epsilon/2}$). ■

We now have the main theorem putting together the entire triple subgroup reduction:

Theorem 15 (triple subgroup reduction)**.** *Let G be a finite group and let H, K be proper subgroups with $N = H \cap K$ normal in G. Then we can compute generalized DFTs with respect to G at the cost of*

- *$|K|/|N|$ many H-DFTs,*
- *$O(|H||K| \log |H/N|/|N|^2)$ many inverse N-DFTs,*
- *$O(|H/N| \log |H/N|)$ many K-DFTs,*

plus $O(|G|^{\omega/2+\epsilon})$ operations, all repeated $O(|G| \log |G|/|HK|)$ many times, for all $\epsilon > 0$.

Proof: By Lemma 13 we can compute the intermediate representation of a G-DFT supported on HK, and applying the map of Lemma 14 to this intermediate representation yields a G-DFT supported on HK. By Theorem 5 we can compute a general G-DFT at the cost of repeating these two steps $O(|G| \log |G|/|HK|)$ many times. ■

4) Triple subgroup structure in finite groups: Our main structural theorem on finite groups is the following

Theorem 16. *There exists a monotone increasing function $f(x) \le 2^{c\sqrt{\log x} \log\log x}$ for a universal constant $c \ge 1$, such that, for every nontrivial finite group G one of the following holds*

1) *G has a (possibly trivial) normal subgroup N and G/N is cyclic of prime order, or*
2) *G has a (possibly trivial) normal subgroup N and G/N has proper subgroups X, Y with $X \cap Y = \{1\}$ and for which $|X||N||Y| \ge |G|/f(G)$.*

To connect this theorem to our usage in the previous sections, think of H as being the subgroup $\overline{X}N$ and K as being the subgroup $N\overline{Y}$, where $\overline{X}$ and $\overline{Y}$ are lifts of X and Y, respectively, from G/N to G.

Proof: Let N be a maximal normal subgroup of G. Then G/N is simple. If it is cyclic of prime order, then we are done. Otherwise we have the following cases, by the Classification Theorem:

1) G/N is an alternating group A_n for $n \ge 5$. In this case, let X be the subgroup of G/N isomorphic to A_{n-1} and Y the trivial subgroup of G/N.
2) G/N is a finite group of Lie Type. In this case, we refer to Table 4, and we have the following description from Carter [21]. For Chevalley and exceptional Chevalley groups, we have that there are subgroups B and U_w^- (for each w in the associated Weyl group W) so that elements of G/N can be expressed *uniquely* as bn_wu_w, where $b \in B$, n_w is a lift of $w \in W$ to G, and $u_w \in U_w^-$ (see Corollary 8.4.4 in Carter [21]). Uniqueness implies that the conjugate subgroup $n_wU_w^-n_w^{-1}$ has trivial intersection with B; also, by an averaging argument, there exists $w \in W$ for which $|Bn_wU_w^-n_w^{-1}| \ge |G/N|/|W|$. We take $X = B$ and $Y = n_wU_w^-n_w^{-1}$. For twisted Chevalley groups, we have an identical situation (see Corollary 13.5.3 in Carter [21]), with subgroup B replaced by B^1 and subgroup U_w^- replaced by $(U_w^-)^1$ (in Carter's notation). Again by an averaging argument there exists $w \in W$ for which $|B^1 n_w (U_w^-)^1 n_w^{-1}| \ge |G/N|/|W|$, and subgroups B^1 and $n_w(U_w^-)^1n_w^{-1}$ have trivial intersection; so we take them as our X and Y, respectively. Finally we verify from Table 4 that in all cases we have $f(|G/N|) \ge |W|$. Thus
$$\begin{aligned}|X||N||Y| &\ge |N||G/N|/|W| \\ &\ge |N||G/N|/f(|G/N|) \ge |G|/f(|G|)\end{aligned}$$
where we used the fact that f is increasing.
3) G/N is a one of the 26 sporadic groups or the Tits group. In this case, we can take $X = Y = \{1\}$, by choosing c in the definition of $f(x)$ sufficiently large.

■

5) Putting it together: Using the structural theorem and the new triple-subgroup reduction recursively, we obtain our final result:

Theorem 17 (main)**.** *For any finite group G, there is an arithmetic algorithm computing generalized DFTs with respect to G, using $O(|G|^{\omega/2+\epsilon})$ operations, for any $\epsilon > 0$.*

Proof: Fix an arbitrary $\epsilon > 0$. Consider the following recursive algorithm to compute a G-DFT. If G is trivial then computing a G-DFT is as well. If G has a proper subgroup H of order larger than $|G|^{1-\epsilon/2}$ then we apply Theorem 7 to compute a G-DFT via several H-DFTs. Otherwise, applying Theorem 16, we obtain a (possibly trivial) normal subgroup N, and two proper subgroups of G, H and K, with $N = H \cap K$. If G/N is cyclic of prime order, we apply Corollary 9 to compute a G-DFT via several N-DFTs. Otherwise, we apply Theorem 15 to compute a G-DFT via several H-DFTs, K-DFTS, and inverse N-DFTs.

Let $T(n)$ denote an upper bound on the operation count of this recursive algorithm for any group of order n. We will prove by induction on n, that there is a universal constant C_ϵ for which
$$T(n) \le C_\epsilon n^{\omega/2+\epsilon} \log n.$$

In the case that we apply Theorem 7, the cost is the cost of $[G : H]$ many H-DFTs plus $A_0[G : H]|G|^{\omega/2+\epsilon/2}$ operations (where A_0 is the constant hidden in the big-oh), and by induction this is at most:
$$\begin{aligned}&C_\epsilon[G : H]|H|^{\omega/2+\epsilon} \log |H| + A_0[G : H]|G|^{\omega/2+\epsilon/2} \\ &\qquad \le C_\epsilon |G|^{\omega/2+\epsilon}(\log |G| - 1) + A_0|G|^{\omega/2+\epsilon}\end{aligned}$$

Name	Family	$\|W\|$	$\|G\|$
Chevalley	$A_\ell(q)$	$(\ell+1)!$	$q^{\Theta(\ell^2)}$
	$B_\ell(q)$	$2^\ell \ell!$	$q^{\Theta(\ell^2)}$
	$C_\ell(q)$	$2^\ell \ell!$	$q^{\Theta(\ell^2)}$
	$D_\ell(q)$	$2^{\ell-1}\ell!$	$q^{\Theta(\ell^2)}$
Exceptional Chevalley	$E_6(q)$	$O(1)$	$q^{\Theta(1)}$
	$E_7(q)$	$O(1)$	$q^{\Theta(1)}$
	$E_8(q)$	$O(1)$	$q^{\Theta(1)}$
	$F_4(q)$	$O(1)$	$q^{\Theta(1)}$
	$G_2(q)$	$O(1)$	$q^{\Theta(1)}$
Steinberg	${}^2A_\ell(q^2)$	$2^{\lceil \ell/2\rceil}\lceil \ell/2\rceil!$	$q^{\Theta(\ell^2)}$
	${}^2D_\ell(q^2)$	$2^{\ell-1}(\ell-1)!$	$q^{\Theta(\ell^2)}$
	${}^2E_6(q^2)$	$O(1)$	$q^{\Theta(1)}$
	${}^3D_4(q^3)$	$O(1)$	$q^{\Theta(1)}$
Suzuki	${}^2B_2(q)$, $q=2^{2n+1}$	$O(1)$	$q^{\Theta(1)}$
Ree	${}^2F_4(q)$, $q=3^{2n+1}$	$O(1)$	$q^{\Theta(1)}$
	${}^2G_2(q)$, $q=3^{2n+1}$	$O(1)$	$q^{\Theta(1)}$

Figure 4. Families of finite groups G of Lie type, together with the size of their associated Weyl group W. These include all simple finite groups other than cyclic groups, the alternating groups, the 26 sporadic groups, and the Tits group. See [18], [20], [21] for sources. The Suzuki, Steinberg and Ree families are also called the *twisted Chevalley* groups.

which is indeed less than $C_\epsilon |G|^{\omega/2+\epsilon}\log|G|$ provided $C_\epsilon \geq A_0$.

In the case that we apply Corollary 9, our cost is p many N-DFTs, plus $A_1|G|^{\omega/2+\epsilon}$ operations, which by induction is at most

$$\begin{aligned} &C_\epsilon p(|G|/p)^{\omega/2+\epsilon}\log(|G|/p) + A_1|G|^{\omega/2+\epsilon} \\ &\qquad \leq C_\epsilon |G|^{\omega/2+\epsilon}(\log|G|-1) + A_1|G|^{\omega/2+\epsilon}, \end{aligned}$$

which is indeed less than $C_\epsilon |G|^{\omega/2+\epsilon}\log|G|$ provided $C_\epsilon \geq A_1$.

Finally, in the case that we apply Theorem 15, let A_2 be the maximum of the constants hidden in the big-ohs in the statement of the Theorem (applied with $\epsilon/2$). Note that by selecting C_ϵ sufficiently large, we may assume that G is sufficiently large, so that two inequalities hold:

$$A_2|H/N|\log|H/N| \leq \frac{|H/N|^{\omega/2+\epsilon}}{4A_2 f(|G|)\log|G|} \quad (11)$$

$$|K/N| \leq \frac{|K/N|^{\omega/2+\epsilon}}{4A_2 f(|G|)\log|G|} \quad (12)$$

and this is possible because Theorem 16 implies that $|H/N|$ (resp. $|K/N|$) are at least $|G|^{\epsilon/2}/f(|G|)$, as otherwise $|K|$ (resp. $|H|$) would exceed $|G|^{1-\epsilon/2}$. Our cost is $|K/N|$ many H-DFTs, $A_2|H||K|/|N|^2\log|H/N|$ many inverse N-DFTs, $A_2|H/N|\log|H/N|$ many K-DFTs, plus $A_2|G|^{\omega/2+\epsilon/2}$ operations, all repeated $A_2|G|\log|G|/|HK| \leq A_2 f(|G|)\log|G|$ times. By induction, this is at most

$$\begin{aligned} &\Big(C_\epsilon|K/N||H|^{\omega/2+\epsilon}\log|H| \\ &\quad + C_\epsilon A_2|H||K|/|N|^2\log|H/N||N|^{\omega/2+\epsilon}\log|N| \\ &\qquad + C_\epsilon A_2|H/N|\log|H/N||K|^{\omega/2+\epsilon}\log|K| \\ &\qquad\qquad + A_2|G|^{\omega/2+\epsilon/2}\Big)\cdot A_2 f(|G|)\log|G| \end{aligned}$$

Now, using Inequalities (11-12), the first three summands are each at most

$$\frac{C_\epsilon|G|^{\omega/2+\epsilon}\log|G|}{4A_2 f(|G|)\log|G|}$$

as is the fourth summand provided $|G|$ is sufficiently large. Thus the entire expression is at most $C_\epsilon|G|^{\omega/2+\epsilon}\log|G|$, as required. This completes the proof. ■

IV. Open problems

Is there a proof of Theorem 16 that does not need the Classification Theorem? A second question is whether the dependence on ω can be removed. Alternatively, can one show that a running time that depends on ω is necessary by showing that an exponent-one DFT for a certain family of groups would imply $\omega = 2$?

Acknowledgements

We thank Jonah Blasiak, Tom Church, and Henry Cohn for useful discussions during an AIM SQuaRE meeting, and Chloe Hsu for her helpful comments on an earlier version of this paper. We thank the FOCS referees for their careful reading and suggestions.

REFERENCES

[1] D. Rockmore, "Some applications of generalized FFTs," in *Proceedings of the 1995 DIMACS Workshop on Groups and Computation*. June, 1997, pp. 329–369.

[2] I. R. Kondor, *Group theoretical methods in machine learning*. Columbia University, 2008.

[3] R. Kondor, "A Fourier space algorithm for solving quadratic assignment problems," in *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 2010, pp. 1017–1028.

[4] D. K. Maslen and D. N. Rockmore, "Generalized FFTs – a survey of some recent results," in *Groups and Computation II*, vol. 28. American Mathematical Soc., 1997, pp. 183–287.

[5] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965.

[6] T. Beth, *Verfahren der schnellen Fourier-Transformation*. Teubner, 1984.

[7] M. Clausen, "Fast generalized Fourier transforms," *Theoretical Computer Science*, vol. 67, no. 1, pp. 55–63, 1989.

[8] U. Baum, "Existence and efficient construction of fast Fourier transforms on supersolvable groups," *computational complexity*, vol. 1, no. 3, pp. 235–256, Sep 1991.

[9] M. Clausen and U. Baum, *Fast Fourier transforms*. Wissenschaftsverlag, 1993.

[10] C. C. Hsu and C. Umans, "A fast generalized DFT for finite groups of Lie type," in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, A. Czumaj, Ed. SIAM, 2018, pp. 1047–1059.

[11] P. Bürgisser, M. Clausen, and M. A. Shokrollahi, *Algebraic Complexity Theory*, ser. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, 1997, vol. 315.

[12] D. K. Maslen, "The efficient computation of Fourier transforms on the symmetric group," *Math. Comput.*, vol. 67, no. 223, pp. 1121–1147, 1998.

[13] D. N. Rockmore, "Fast Fourier transforms for wreath products," *Applied and Computational Harmonic Analysis*, vol. 2, no. 3, pp. 279 – 292, 1995.

[14] D. Maslen and D. Rockmore, "Separation of variables and the computation of Fourier transforms on finite groups, I," *Journal of the American Mathematical Society*, vol. 10, no. 1, pp. 169–214, 1997.

[15] D. N. Rockmore, "Recent progress and applications in group FFTs," in *Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on*, vol. 1. IEEE, 2002, pp. 773–777.

[16] D. Maslen, D. N. Rockmore, and S. Wolff, "The efficient computation of Fourier transforms on semisimple algebras," *Journal of Fourier Analysis and Applications*, vol. 5, no. 24, pp. 1377–1400, 2018.

[17] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[18] A. Lev, "On large subgroups of finite groups," *Journal of Algebra*, vol. 152, no. 2, pp. 434–438, 1992.

[19] C. C. Hsu and C. Umans, "A new algorithm for fast generalized DFTs," *CoRR*, vol. abs/1707.00349v3, 2018, full version of [10].

[20] Wikipedia, "List of finite simple groups — Wikipedia, the free encyclopedia," 2017, [Online; accessed 30-June-2017].

[21] R. W. Carter, *Simple groups of Lie type*. John Wiley & Sons, 1989, vol. 22.